ALTERATION OF TRANSCRIPTION BY NON-CODING ELEMENTS IN THE HUMAN GENOME

A Dissertation
Presented to
The Academic Faculty

By

Andrew Conley

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biology

Georgia Institute of Technology
August 2012

ALTERATION OF TRANSCRIPTION BY NON-CODING ELEMENTS IN THE HUMAN GENOME

Approved by:

Dr. I. King Jordan, Advisor
School of Biology
Georgia Institute of Technology

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Dr. Leonardo Mariño-Ramírez
National Center for Biotechnology Information
National Institutes of Health

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Dr. Roger Wartell
School of Biology
Georgia Institute of Technology

Date Approved: June 25, 2012

To my wife Kathryn and my daughter Charlotte…

ACKNOWLEDGEMENTS

I must first thank my advisor, Dr. Jordan. He has been an extraordinary teacher and guide throughout my time as his PhD student. I have been extremely fortunate to have his guidance over the years and I have learned most of what I know about the scientific process from him. I am also thankful to my committee members, Jung Choi, Leonardo Mariño-Ramírez, John McDonald, and Roger Wartell, for their help in my research and academic career.

Apart from Dr. Jordan himself, the people in the Jordan lab, both past and present members, have been extraordinary in my time as a PhD student. Jittima Piriyapongsa and Ahsan Huda were extraordinarily helpful to me in my early years, while Jianrong Wang has been extremely helpful in the later years. Lee Katz, Eishita Tyagi and Daudi Jjingo have been wonderful friends over the past few years.

I am very grateful to my family and all of the support and encouragement that I have received from them. Both of my parents, Jim and Colleen Conley, have instrumental in my education, particularly my mother for encouraging my interest in biology. My mother-in-law Jan and father-in-law David have been extraordinarily encouraging throughout my education, and I wish to thank them as well.

More than anyone else, I would like to thank my wife wonderful wife Kathryn. She has been nothing but loving, supportive and understanding throughout a very long PhD process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF TERMS

CAGE	Cap Analysis of Gene Expression
ERV	Endogenous Retrovirus
H3K4Me1	Histone H3 lysine 4 mono-methylation
H3K4Me2	Histone H3 lysine 4 di-methylation
H3K4Me3	Histone H3 lysine 4 tri-methylation
H3K9Ac	Histone H3 lysine 9 acetylation
H3K9Me1	Histone H3 lysine 9 mono-methylation
H3K9Me3	Histone H3 lysine 9 tri-methylation
H3K27Ac	Histone H3 lysine 27 acetylation
H3K27Me3	Histone H3 lysine 27 trimethylation
H3K36Me3	Histone H3 lysine 36 trimethylation
IAP	Intracisternal A Particle
LTR	Long Terminal Repeat
MaLR	Mammalian LTR Retrotransposon
MIR	Mammalian Interspersed Repeat
PET	Paired-End diTag
RNA PolII	RNA Polymerase II
TE	Transposable Element
TE-TSS	TE-derived Transcription Start Site
TE-TTS	TE-derived Transcription Termination Site

SUMMARY

The human genome contains ~1.5% coding sequence, with the remaining 98.5% being non-coding [1]. The functional potential of the majority of this non-coding sequence remains unknown. Much of this non-coding sequence is derived from transposable element (TE) sequences. These TE sequences contain their own regulatory information, e.g. promoter and transcription factor binding sites. Given the large number of these sequences, over 4 million in the human genome, it would be expected that the regulatory information that they contain would affect the expression of nearby genes. This dissertation describes research that characterizes that alternation of and contribution to the human transcriptome by non-coding elements, including TE sequences.

Research advance 1: Chapter 2 evaluates the abundance of cis-natural antisense transcript (cis-NAT) promoters derived from TE insertions within human genic loci. TE sequences require their own promoters for transcription, and previous examples of TE-derived promoters are known. Here it is shown that TE sequences provide cis-NAT promoters inside of human genes and that these TE-derived cis-NAT promoters are more common toward the 3'-end of genic loci.

Research advance 2: In chapter 3, the presence of alternative promoters for human genes derived from endogenous retrovirus (ERV) insertions is investigated genome-wide. Paired-end diTag (PET) and Cap Analysis of Gene Expression (CAGE) data from mammary epithelial cell lines are used to identify transcripts of human genes that are initiated from ERV sequences. The work shown here demonstrates that ERV sequences have contributed promoters to over 100 different human genes.

Research advance 3: In chapter 4, we review techniques for characterizing transcription factor binding sites derived TE sequences using ChIP-seq data. TE sequences contain transcription factor binding sites (TFBS) that they use for their own transcription, as well as potentially harboring sequences similar to other TFBS. The spread of TE sequences in a host genome could thus create additional TFBS [2]. Such novel TFBS could greatly affect the expression of nearby host genes. However, the short read length of ChIP-seq data necessitates special care when characterizing DNA-protein interactions involving TE sequences. The methods shown here are generally applicable for characterizing TFBS derived from TE sequences.

Research advance 4: In Chapter 5, the regulation of cis-NAT promoter activity via chromatin modification is characterized. The activity of cis-NAT promoters is shown to be correlated with the local presence of activating histone modifications, *e.g.* H3K9Ac, and anti-correlated with the presence of the repressive H3K27Me3 modification [3, 4]. It is also shown that the distribution of histone modifications near promoters is very similar to that of canonical promoters for protein coding genes. Finally, the cis-NAT promoters characterized are shown to be active in fewer cell types on average than promoters of protein coding genes. The regulation of cis-NAT promoters via chromatin modification is indicative of the function of cis-NAT promoters in the human genome.

Research advance 5: Chapter 6 reviews instances in which ERV sequences have been shown to have an effect on the host transcription via epigenetic modification of the ERV sequence. The control of ERV sequences via repressive epigenetic modifications has the potential to lower the expression of host genes near the ERV sequences. Several studies reviewed demonstrate such effects of ERV sequences on the expression of nearby

genes. Studies showing the converse, examples of ERVs bearing activating epigenetic marks positively affecting gene expression, are also reviewed. The studies reviewed here demonstrate that the epigenetic modification of ERV sequences has the potential to effect host gene expression.

Research advance 6: In chapter 7, the extent to which TE sequences within gene loci terminate the transcription of human genes, and the cell type-specificity of such termination, are explored. TE sequences within gene loci are more often than not found in the antisense orientation, likely resulting from selection against sense insertions that could terminate gene transcription. Utilizing high-throughput PET data, many thousands of alternative transcription termination sites (TTS) derived from TE sequences (TE-TTS) are characterized across eight human cell types. The relative strengths of TTS derived from sense and antisense sequences as well as different TE families are evaluated; TE-TTS in the sense orientation are generally much stronger, and the different TE families show substantial differences in the strength of TE-TTS derived from them. We further show that TE-TTS provide many highly cell type specific TTS. These results demonstrate that TE sequences have had a major effect on the expression of human genes via the termination of transcription.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Eukaryotic genomes contain abundant non-coding DNA

Transposable elements (TE) are DNA sequences which have the ability to 'transpose' or move about in the genome. There are four main types of TEs: SINEs, LINEs, LTR elements and DNA transposons. The first three types move through in the genome via retrotransposition where the element is first transcribed into mRNA then reverse transcribed and into the genome. The fourth type moves by a 'cut-and-paste' mechanism where the actual DNA sequence of the element is excised from the genome and inserted elsewhere. SINEs (short interspersed nuclear elements) are small TE sequences, typically derived from non-coding RNAs, *e.g.* tunas, 7SL RNAs, etc. SINEs do not encode any of the enzymes required for retrotransposition and instead rely on enzymes encoded from other classes of TEs. LINEs (long interspersed nuclear elements) are a family of TEs which do encode enzymes necessary for retrotransposition; it is thought that SINEs typically rely on LINE-encoded enzymes for retrotransposition. LTR elements are a class of TEs which have, on either end of their internal sequences, long terminal repeats, for which LTR elements are named. These LTRs are direct repeats, *i.e.* the 5' LTR and 3' LTR are identical. Like LINEs, these TEs encode the enzymes and other proteins required for their retrotransposition, though the process is substantially different from LINEs.

Transposable elements (TEs) often make up a substantial fraction of non-coding DNA in mammalian genomes. The human genome is nearly 50% annotated TE

sequences [5]. However, a recent study suggests that much of the of remainder of the human genome is composed of highly diverged TE sequences, and that the human genome may be as much as 69% TE sequence [6]. An example of the abundance of TE-derived sequence in the human genome is shown in Figure 1.1. Three human protein coding genes (black) are present in this short region, however they are dwarfed by the number of TE sequences (grey). This region is fairly representative of the human genome as a whole; the amount of sequence coding for human genes is vastly exceeded by TE sequences.

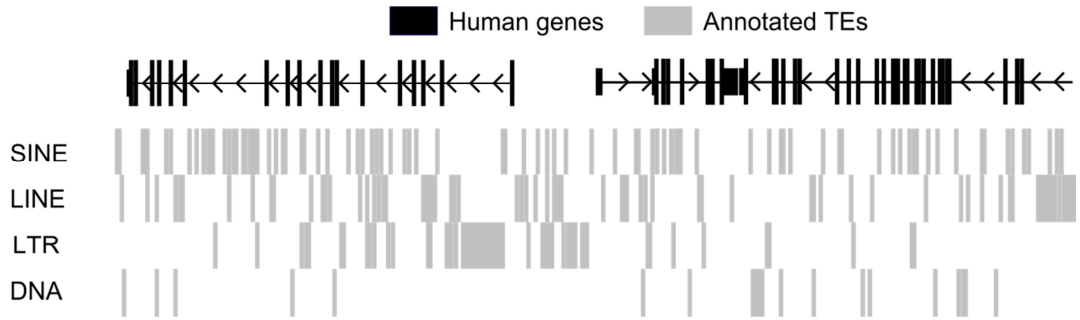


Figure 1.1. The human genome is dense with TE sequences. Three human genes (black) are present in the region. Various TE sequences within the region from the four TE classes are shown in grey.

In contrast to other species, there are relatively few actively transposing TE families in the human genome [1, 7, 8]. Indeed, only a few specific subfamilies of the Alu, L1 and SVA families are currently active in the human genome [1]. Alu elements are an active family of SINEs specific to the primate lineage, originating from the expansion of a7SL RNA early in primate evolution. The modern family of Alu elements

has been extremely successful, expanding to ~1.2 million copies in the human genome [5, 9, 10]. L1 elements are an active family of LINEs found throughout mammals, and contribute 17% of the human genome [1, 5]. SVAs are a relatively recent family of non-coding TEs derived from a fusion of other TE families. SVAs are active, but rare; only ~4,500 copies are present in the human genome [11]. In addition to these families, there are many thousands of sequences derived from functionally dead TE families in the human genome. Mammalian Interspersed Repeats (MIRs) and L2s are tRNA derived SINEs and LINEs, respectively. Though both families stopped spreading some time ago, many more of these sequences than expected have been conserved over evolutionary time, suggesting that they have exapted to play some role for the host genome [12]. Importantly, these sequences contain regulatory sequences to promote their own transcription and spread. L1 elements contain, for example, their own internal promoter and terminator sequences [13, 14]. When inserted within or near to a host gene, these regulatory sequences could drastically alter the transcription of the host gene.

Active TE Sequences can generate many additional copies in a relatively short evolutionary time. There are, for example, many TE sequences which inserted after the divergence between human and chimpanzee. Between the human and chimpanzee genomes, there are over 9,000 TE insertions lacking in one genome or the other, and an additional ~4,000 L1 insertions. Interestingly, while ERVs appear to be dead or near dead in the human genome, several families have multiplied in the chimpanzee genome yielding several hundred new insertions [1, 8].

The human-chimpanzee divergence was very recent in evolutionary terms. The differences in TE sequences are much greater over a longer evolutionary time. Human

and mouse have very different sets of active TEs. While the human genome contains ~1.2 million Alu sequences, the Alu family is not present in the rodent lineage [1, 7]. While LTR elements are functionally dead in humans, they are the most active TEs in mouse, particularly the intracisternal-A particle (IAP) family of ERVs. An example of the dramatic differences in TE content between human and mouse is shown for the CHRNA2 locus (Figure 1.2). It can be seen that the coding exons of the CHRNA2 gene (thick black blocks) are present in both species, i.e. the coding exons are located within aligned regions (blue blocks). It is worth noting that the annotated promoters from human and mouse do not arise from homologous loci (see below). For both the human and mouse genomes, there are many large gaps in the alignment representing sequences not present in the other species; many of these gaps are due solely to the presence of lineage-specific TE sequences. For example, the highlighted TE sequence (red) resulted from the insertion of a B2 element, a rodent-specific family of SINEs, and it leads to a gap in the alignment as there is no homologous sequence in the human genome. Being lineage-specific, these TE sequences and whatever effect they have on the transcription of host genes are also lineage-specific. The vast number of differences in the TE sequence content between human and mouse, much of it within coding loci, virtually guarantees that TE sequences have altered gene expression via non-coding means. Indeed, many such examples are known, and five of the works in this dissertation add to the known effects of TE sequences on gene expression.

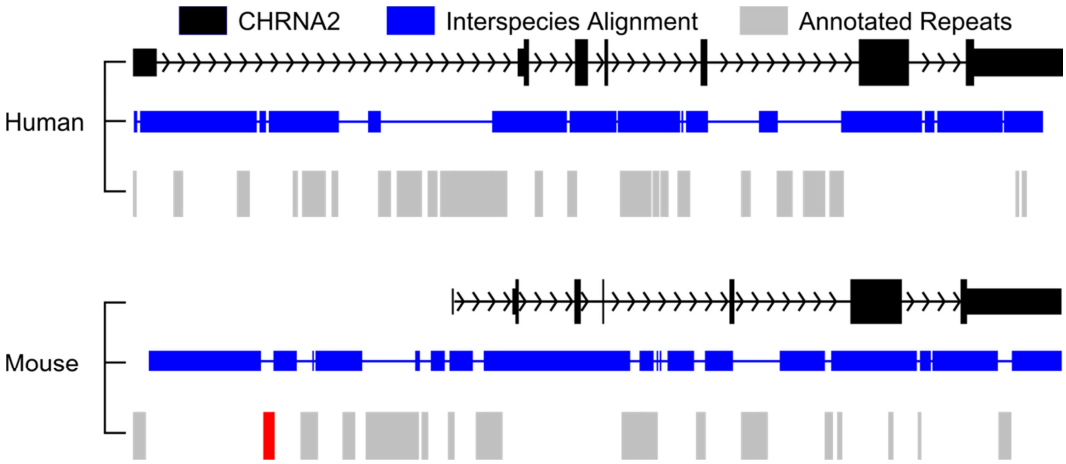


Figure 1.2. Many TE sequences at the CHRNA2 locus are lineage-specific. Homologous annotated CHRNA2 transcripts (black) from human and mouse are shown above regions which align between the two genomes (blue blocks) and annotated TE sequences within each genome (grey). A B2 insertion (red) in the rodent lineage generated a rodent-specific genomic sequence resulting in a gap in the alignment with the human genome. There are many other lineage-specific TE sequences (grey) that correspond to gaps in the human-mouse genome sequence alignment.

The human genome contains many functional non-coding elements

There exists in the human genome a wide variety and a very large number of non-coding elements that influence gene transcription. The mechanisms by which these non-coding elements influence transcription are equally varied. This dissertation is concerned with three of the most prominent varieties of non-coding regulatory elements: promoters, transcription termination sites (TTS), and transcription factor binding sites (TFBS). Of interest, it has been show that various TE sequences have been exapted to function as all of these non-coding elements.

Every gene has some non-coding elements which influence its transcription: at least one promoter and at least one terminator. Promoters are where the transcription of the mRNA begins, and can broadly active or cell type-specific, and alternative promoters

allow for a condition and cell type-specific expression of genes and isoforms [15-17]. It is worth noting that alternative promoters derived from TE sequences have been previously described [18]. Promoters, however, are not exclusive to protein-coding genes; there are many known, long non-coding transcripts in the human genome, many of which show cell type-specific expression as protein coding genes do [19]. At the opposite end of the transcript is the polyadenylation signal and transcription termination site. As with promoters, human transcription terminators derived from TE sequences have been known for some time [20, 21]. Promoters can substantially alter the amount of mRNA produced, while both alternative promoters and terminators can alter the coding sequence, potentially changing both the quality and quantity of the mRNA.

The third category of concern here, transcription factor binding sites, including enhancers, alter the expression of genes not via initiation or termination of transcription, but by recruiting proteins required for transcription to the promoter proper. As with promoters and terminators, transcription factor binding sites and enhancers derived from TE sequences have been described. For example, a very old SINE insertion was shown to act as an important enhancer in neural development [22, 23]. Recently, it was shown that a family of LTR elements, MER41, has greatly expanded number of functional STAT1 binding sites in the human genome [24].

High-throughput techniques and massively-parallel sequencing and have drastically altered the study of genome function

Previous methods of characterizing genomic function fell short in a number of ways, particularly with regard to the function of non-coding elements derived from TE

sequences. EST sequencing using Sanger sequencing is far too low throughput to capture the breadth and variation in human gene expression or to accurately characterized DNA-protein interactions genome wide. The large size of the human genome means also that genome-wide (tiling) microarrays are unfeasible. Further, microarray based techniques cannot detect TE-derived sites due to their repetitive nature. Thus TE sequences and any functional elements they may provide cannot be detected via such methods, *e.g.* the pilot phase of the ENCODE project [25]. Current technologies using massively-parallel sequencing allow for the genome-wide, unbiased characterization of RNA transcripts (RNA-seq) and DNA-protein interactions (Chromatin Immunoprecipitation followed by high-throughput sequencing, ChIP-seq). For example, an early large-scale study using RNA-seq and the Illumina platform showed that nearly all multi-exon human genes are alternatively spliced in some tissue, far more than previously appreciated [26]. Another early study used the Illumina platform to characterize NRSF binding genome wide via ChIP-seq [27]. Several of the works shown here make use of ChIP-seq data and/or RNA-seq data from the production phase of the ENCODE project to characterize non-coding functional elements in the human genome [28, 29].

Along with massively parallel sequencing technologies, several molecular techniques have greatly aided the characterization of promoters and terminators. The first, cap analysis of gene expression (CAGE), generates short tags from the 5' end of mature mRNA transcripts, which, when sequenced and mapped to the genome, can identify TSS and promoters [30, 31]. A second technique, paired-end diTag (PET) generates similar data, but allows for the characterization of both the 5' and 3' ends of a transcript. Combining these techniques with massively parallel sequencing allows for the

in-depth interrogation of promoter and terminator activity. Importantly, as these techniques are not array based, they allow for the functional characterization of TE-derived sequences. Indeed, CAGE has been previously been combined with massively-parallel sequencing to characterize the transcription of TE sequences across a large number of human and mouse cell types [32]. Data generated using the CAGE and PET technique are used in the majority of the studies shown here to characterize the contribution of non-coding sequences to human transcription.

Antisense transcription is pervasive in the human genome

One of the more interesting observations in recent years is that the large majority of the human genome, including both strands, is transcribed at some point in time.
A substantial fraction of this transcription is in the form of non-coding cis-natural antisense transcripts (cis-NATs) [33, 34]. These transcripts are transcribed from the opposite strand of a gene coding locus, and would thus be antisense and complementary to the sense product. Several examples of cis-NAT transcription negatively affecting sense product abundance are known [35, 36]. Whether or not general cis-NAT transcription is functional, or simply noise, remains to be seen. The sources of cis-NAT transcription and regulation are characterized in two of the studies presented here.

Alternative transcription termination has attracted recently attracted great interest

It has previously been estimated that at least 50% of human genes contain alternative termination sites; utilization of these sites by an elongating RNA polymerase results in different transcript isoforms [37, 38]. The effect of such alternative termination on gene expression goes beyond the transcriptional level, effecting mRNA lifespan and the final

translated protein product. Utilizing intronic termination sites, genes encoding receptor tyrosine kinases have been found to produce shorter transcripts encoding truncated receptors missing transmembrane domains. Proteins produced from these truncated transcripts may act as molecular decoys, sequestering other proteins that would otherwise interact with the full-length products. Recent studies have shown that cancer cells and other proliferating cells broadly express transcripts with shortened 3'UTRs compared to differentiated cells [39, 40]. Conversely, it was shown that cells going through differentiation progressively express transcripts with longer 3'UTRs [41]. These studies strongly suggest that the use of the use of alternative termination sites to generate shorter or longer transcripts is an important part of gene regulation. In this dissertation, the cell type-specific use of alternative termination sites derived from TE sequences is explored.

CHAPTER 2

HUMAN CIS-NATURAL ANTISENSE TRANSCRIPTS INITIATED BY TRANSPOSABLE ELEMENTS

Abstract

The capacity of human transposable elements (TEs) to promote cis natural antisense transcripts (cis-NATs) is revealed by the discovery of 48,718 human gene antisense transcriptional start sites (TSSs) within TE sequences. TSSs that yield cis-NATs are overrepresented among TE sequences, and TE initiated cis-NATs are more abundant close to the 3' ends of genes. The TE sequences that promote antisense transcription within human genes are relatively ancient suggesting that selection has acted to conserve their function.

Introduction

Cis natural antisense transcripts (cis-NATs) are RNAs that are transcribed from the antisense strand of a gene locus, which are thus complementary to the RNA transcribed from the sense strand. It is becoming increasingly apparent that cis-NATs are used to regulate the expression of human genes [33, 42, 43]. Cis-NATs may regulate expression at the transcriptional level, via the avoidance of transcriptional collisions [44], or post-transcriptionally through any one of the number of double stranded RNA (dsRNA) induced regulatory pathways collectively known as RNA interference [45].

Transposable elements (TEs) have been shown to contribute a variety of non-coding RNAs that act as dsRNA regulators of gene expression across diverse eukaryotic

species including human. Indeed, TEs encode a number of distinct classes of regulatory RNAs including short interfering RNAs [46-48], microRNAs [49, 50], repeat-associated small interfering RNAs [51] and piwi-interacting RNAs [52]. It appears that multiple distinct RNA interference mechanisms have evolved independently as genome defense mechanisms against TEs only to be later coopted to regulate host genes [49]. There are three reasons to believe that TEs may represent a potentially rich source of cis-NATs that can regulate human gene expression: i-the abundance of TEs in the human genome [1], ii-the ability of TE sequences to promote transcription [53] and iii-the relationship between TEs and RNA interference. To explore this possibility, we have conducted a genome-scale survey of the ability of TE sequences to contribute cis-NATs to human genes.

Methods

Identification of TE-TSSs from CAGE data

A library of 1,551,672 human CAGE sequence tags [31, 54] was download from the Japanese National Institute of Genetics website (http://genomenetwork.nig.ac.jp/public/download/cage_Database_e.html). The data used in the manuscript correspond to the 2007.3.28 release. Human CAGE sequence tags were mapped to the hg18 version, *i.e.* the National Center for Biotechnology Information release 36, of the reference human genome sequence as previously described [30]. A browser extensible data (BED) format custom track with all CAGE tag-to-genome mapping coordinates, available on request, was generated in order to integrate the CAGE data with the human genome reference sequence annotations available from the UCSC Genome Browser Database [55]. The UCSC Table Browser [55] was used together with

a series of custom developed Perl scripts, available on request, to identify the intersection between human transcriptional start sites (TSSs) identified by CAGE tags with human transposable elements (TEs). To identify TE-derived TSSs, the human CAGE custom track was intersected with the Repeat Masker [5] (rmsk track) annotation track using 100% overlap and non TE-classes of repetitive DNA were subsequently eliminated from consideration. Specific TE family/class identities of the resulting TE-TSSs were determined by mapping these results back to the rmsk track and parsing the annotation therein. The observed percentages of 1) all, 2) sense and 3) antisense TE-TSS were determined for seven individual classes/families of TEs. The observed values were compared to the expected values that were determined by calculating their relative frequencies in the RepeatMasker annotation for the whole genome. Observed and expected values sum to 100% over all TE categories.

Human genes and TSSs from CAGE data

The UCSC Genome Browser 'Old Known Genes' track annotations (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=99200641&c=chr7&g=knownGeneOld2) were used to define the coordinates of human protein-coding genes on the hg18 reference sequence. These human gene annotations were chosen because they represent a conservative set of gene definitions that are supported by multiple lines of evidence from the SWISS-PROT, TrEMBL and Genbank databases [56, 57]. A custom Perl script was used to divide all human genes, from their 5' to 3' ends, into twenty equal sized bins, and the TSSs identified from CAGE data were mapped into gene-specific bins. Where the Old Known Genes track annotates multiple alternative transcript variants transcribed from a single

genomic locus in the same direction, the resulting TSSs locations were only counted once. The antisense-versus-sense orientations (ratios) of TSSs were then considered with respect to their location in each bin along the gene lengths. This procedure was repeated for 1) non TE-TSS, 2) all TE-TSS and 3) individual families/classes of TE-TSS.

Relative ages of TE-TSSs

The relative ages of different families/classes of TEs were taken from the RepeatMasker analysis of the human genome reference sequence [1]. Since TE sequences in the human genome are derived from, and related to, copies of once active elements, and have subsequently accumulated mutations after insertion in the genome, the elements can be clustered into phylogenetic trees and grouped into related families/classes. The ensemble of sequences in any given class can be used to compute a consensus sequence, which is taken to represent the ancient (active) copy of the TE [58]. Such consensus sequences have been extensively constructed from human genome TEs and are available in the Repbase database [59, 60]. Ages of TEs can be then be inferred by comparing the sequence divergence between the extant element sequence identified in the genome and its most closely related consensus sequence [61]. This information is made available as the 'millidiv' output, *i.e.* number of substitutions per 1,000 sites, from the RepeatMasker program. Percent substitution of extant TEs from consensus sequences was used to show that the human genome has experienced successive waves of expansion of different families/classes. Consequently, some families/classes are substantially older (or younger) than others. The most ancient families in the human genome are the L2 and MIR families, while the youngest are L1 and Alu [1]. These are the specific findings that are used to consider the relative ages of TE-TSSs derived from different families/classes.

Divergence (*d*) of extant TE sequences identified in the human genome from their consensus sequences were also used to evaluate the relative ages of TEs within the Alu family of elements. To do this, individual Alu insertion millidiv values were converted to Jukes-Cantor DNA sequence distances [62] using the following formula:

$$d = -3/4 * \ln[1 - 4/3(millidiv\ /1{,}000)$$

Then, the average and standard deviation *d*-values were computed and compared for Alu elements that donate TSSs versus those that do not using the Student's t-test.

Human-Mouse conservation of TSS

To evaluate the relative human-to-mouse evolutionary conservation, *i.e.* presence/absence of orthologous insertions, of 1) non TE-TSS, 2) all TEs and 3) all TE-TSS, the UCSC Genome Browser 'liftOver' utility was run locally. This program allows for annotation coordinates from one genome, or build, to be directly transferred to a second genome based on where they correspond. In the case of the human-mouse comparison, the coordinate correspondence is based on whole genome sequence alignments [63] represented in the Mouse Chain track (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=99200641&c=chr7&g=chainMm9). To count the number of base pairs conserved between human and mouse for the different categories mentioned above, the 'Base Coverage' utility of the Galaxy Server [64] was used. Relative conservation was measured as the fraction of base pairs conserved for the different categories.

Results and Discussion

Genome-scale identification of cis-NATs

In order to evaluate the capacity of TEs to contribute cis-NATs to the human genome, we took advantage of a relatively new technology – cap analysis of gene expression (CAGE) [30] – to define the location of transcriptional start sites (TSSs) in the human genome. CAGE relies on the isolation of full length cDNAs using biotinylated mRNA caps. Linkers are ligated to the 5' ends of the full length cDNAs, and the first 20bp of the cDNAs are cleaved with restriction enzymes. The resulting fragments are then amplified, concatamerized and sequenced allowing for the high-throughput characterization of the 5' ends of mRNAs. Mapping of the 5' mRNA end CAGE sequence tags to the genome unambiguously identifies TSSs. Then, the location and orientation of human TSSs can be compared to gene and TE annotations to assess the relationship between cis antisense transcription and TEs.

A library of $>1.5x10^6$ human CAGE sequence tags are available for download from the Japanese National Institute of Genetics website. We mapped these CAGE sequence tags to transcriptional units (TUs) in the human genome, and compared their locations to those of TEs, to assess whether TEs provide cis-NATs. A TU is defined here as a single protein-coding locus, with a characteristic strand orientation, bounded by the most 5' and 3' transcription start and termination sites respectively. The UCSC Genome Browser database [55] KnownGenes annotations were used to locate TUs on the human genome sequence. KnownGenes annotations were chosen because they are supported by multiple sources of information including SWISS-PROT, TREMBL and Genbank

mRNAs. A single TU may cover alternative transcript variants in the same orientation, and more than one TU may overlap at a single genomic locus.

A total of 869,085 CAGE sequence tags were mapped to 39,288 human genome TUs. The majority (639,490 or ~74%) of human TU-TSSs defined in this way correspond to sense transcripts, *i.e.* in the same orientation of the protein-coding mRNA. On average, each human TU has 16.3 sense TSSs. The prevalence of sense oriented TSSs in the human genome is consistent with previous results and reflects the initiation of mRNA transcripts from multiple alternative promoters [54, 65]. The relative excess of sense oriented TSSs is also thought to be due to selection against initiation of antisense transcription based on avoidance of collisions between the RNA transcription machinery tracking along the DNA [44]. Human TUs also have numerous anti-sense oriented TSSs (229,595 or ~26%), which correspond to cis-NATs. The abundance of human cis-NATs is underscored by the fact that the average human TU has 5.8 antisense TSSs.

TEs initiate antisense transcription

The location of TEs in human TUs were taken from the RepeatMasker [5] annotation of the genome, and these data were used to discover transcripts that are initiated within TEs inserted into human TUs. 176,578 (~20%) of human TU-TSSs were found to be initiated from within TE sequences. These data underscore the substantial capacity of TEs to promote transcription in human gene regions. TSSs that are initiated from within TEs are more likely to be found in the antisense orientation than non-TE-TSSs. 48,718 of TE-TSS are found in the antisense orientation, yielding a TE antisense/sense ratio of 0.38 compared to 0.35 for non-TE-TSSs. This difference, while moderate, is highly significant using a χ^2 analysis of 2x2 contingency table

(48,718/127.860 non-TE-TSS antisense/sense=180,877/511,630 χ^2=156.6 *P*=6e-36). In other words, the TSSs that originate from within TE sequences are significantly enriched for cis-NATs, and this observation can not be explained by random sampling alone. The vast majority of TEs that encode TSSs are found in introns (98.2%), which is consistent with the transcriptional collision model [44] for their mechanism of regulatory action since these cis-NATs may not necessarily form dsRNA with mature sense transcripts.

The enrichment of cis-NATs initiated from within TE sequences is even more marked when the distribution of TSSs across TUs, from the 5' to the 3' ends, is observed. Human TUs were divided into 20 equal sized bins and antisense/sense TSS ratios were calculated for each bin. When bin-specific averages across all human TUs are plotted, the ratio of antisense/sense TE-TSSs increases progressively from the 5' to the 3' ends of human TUs (Figure 2.1). The slope of this trend is positive, and the correlation is statistically significant. This enrichment suggests the possibility that antisense TE-TSSs near the 3' ends of genes are more efficacious regulators and thus favored by selection. Under the transcriptional collision model [44], the preponderance of cis-NATs initiated near the 3' ends of genes would provide for more opportunities for collisions between RNA polymerase complexes tracking along opposite strands of the DNA and less chance for sense transcription complexes to get through to the ends of the genes. The opposite 5'-to-3' trend in antisense/sense TSS ratios is seen for non-TE-TSSs. There is a slight decrease in the antisense/sense ratio along human TUs; although, this trend is far less pronounced and not statistically significant (Figure 2.1).

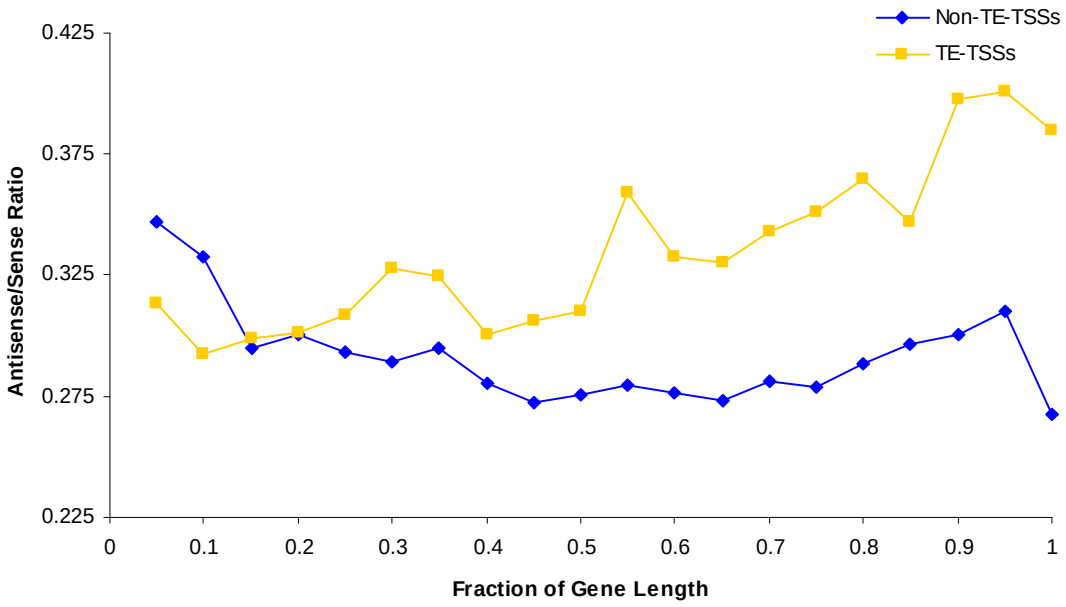


Figure 2.1. Ratio of antisense/sense TSSs along human genes. Human TUs were divided into twenty equal size bins, and ratios of the numbers of antisense/sense TSSs were calculated for each bin. Average bin-specific ratios are shown for TSSs initiated within TEs (TE-+ in grey) and TSSs not initiated from TEs (Non-TE TSSs in black). Linear regression was used to plot the slope of the antisense/sense ratio trend along bins from 5'-to-3' gene ends and the Spearman rank correlation coefficient (R) was used to evaluate the significance of the trends. For TE-TSSs y=10e-2, R=0.87, t=7.62, P=5e-7. For non TE-TSSs y=-3e-2, R=-0.34, t=1.54, P=0.14.

The relative excess of antisense transcripts initiated from TEs and their enrichment closer to the 3' ends of TUs suggests the possibility that they may yield cis-NATs with biologically significant regulatory activities. If this is indeed the case, then one may expect natural selection to preserve these functionally active TE-derived cis-NATs. Accordingly, TEs that initiate cis-NATs are predicted to be older than those that do not initiate transcription owing to the fact that they have been preserved in the genome by selection. The age distribution of the TEs that donate cis-NATs was analyzed to evaluate this prediction. The observed proportions of elements from different TE

classes/families that donate cis-NATs were compared to the expected proportions based on their relative frequencies in the genome. Consistent with the expectation, relatively ancient elements are significantly overrepresented (Figure 2.2). For instance, there are more TU-TSSs derived from the ancient L2 and MIR families than expected by their genome frequencies. Members of younger element families, such as L1 and Alu, initiate significantly fewer TU-TSSs than expected based on their genome frequencies.

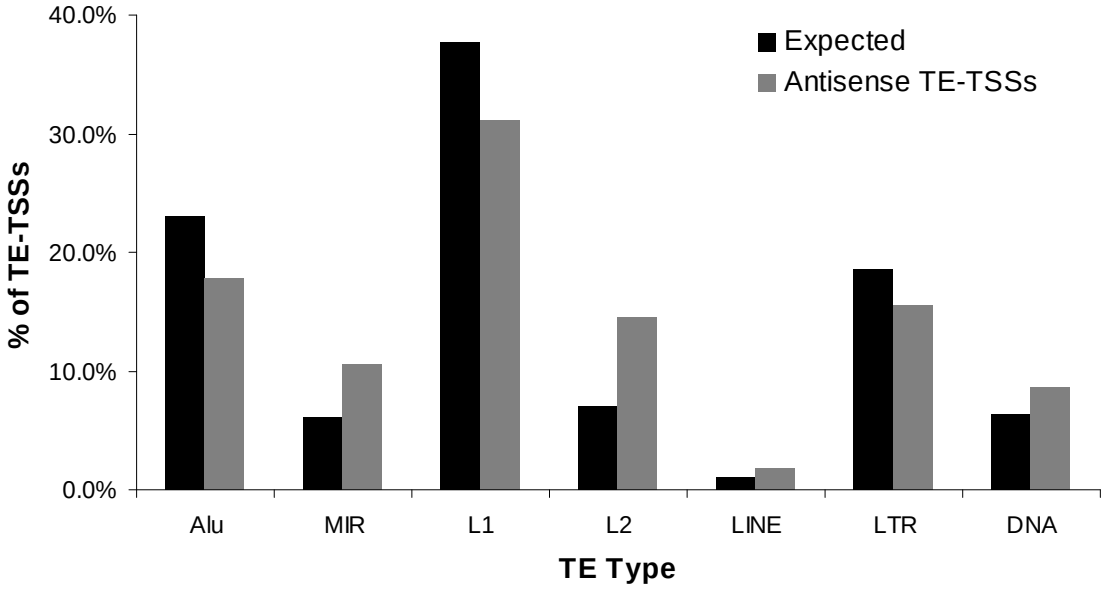


Figure 2.2. Relative proportions of TE-derived cis-NATs. Human genome TEs are broken down into seven classes/families and the relative percentages of TE-derived cis-NATs are shown in gray for each family. The expected percentages of TE-derived cis-NATs, based on the genome proportions of each class/family, are shown in black. A $\chi 2$ test for goodness of fit was used to compare the observed versus expected proportions of TE-derived cis-NATs. The differences between the observed and expected distributions across classes are highly statistically significant ($\chi 2$=7,671 P=0).

The relative ages of TEs within families can be measured by taking the divergence of the TE sequences from their family consensus sequence, since older TEs have accumulated more substitutions, on average, than younger elements [1]. For the

younger TE families, relative divergence values indicate that the TEs that donate antisense TSSs are older than TEs from the same family that do not donate TU-TSSs. For instance, Alu elements that initiate antisense transcription with human TUs have significantly greater sequence divergence from their consensus sequences than Alus that do not co-locate with TSSs (average±standard deviation Jukes-Cantor distance for Alu-antisense-TSS=0.15±0.05, Alu-non-TSS=0.14±0.05, t=19.42, *P*=6e-84). This suggests that many antisense TE-TSSs are in fact conserved by selection and also helps to resolve a standing question as to why older elements of some classes of TEs, such as Alus, are enriched in gene regions. Alus insert more frequently into AT-rich DNA but are preferentially retained in GC-rich gene regions; this has been taken to suggest that they are conserved in gene regions by virtue of some unknown functional role that they play for those genes [1]. Our data indicate that, in the case of some Alu sequences, their functional role is related to the initiation of regulatory cis-NATs.

Another way to evaluate the relative ages of TEs is to compare their evolutionary conservation based on presence/absence patterns of orthologous insertions between related species. Using this approach, we compared the human-to-mouse conservation of TEs that encode TSSs versus those that do not. 15.3% percent of TE-TSSs are conserved between human and mouse versus 2.8% percent of TEs that do not encode TSSs; this difference is statistically significant (χ^2=4x106 *P*=0). The greater between species conservation of TE-TSSs is further evidence consistent with the action of purifying selection based on function.

Conclusions

The ability of TEs to contribute regulatory sequences to eukaryotic genomes was discovered through a number of case-by-case studies on individual genes [66, 67]. Later, genome-scale approaches began to uncover just how widespread this phenomenon is, particularly in mammalian genomes with high TE copy numbers [68-70]. Relying on a genome-scale approach for the identification of TSSs, we have shown that TEs contribute tens-of-thousands of cis-NATs to human genes. The potential regulatory effects of these TE derived antisense transcripts are substantial.

CHAPTER 3

RETROVIRAL PROMOTERS IN THE HUMAN GENOME

Abstract

Endogenous retrovirus (ERV) elements have been shown to contribute promoter sequences that can initiate transcription of adjacent human genes. However, the extent to which retroviral sequences initiate transcription within the human genome is currently unknown. We analyzed genome sequence and high-throughput expression data to systematically evaluate the presence of retroviral promoters in the human genome. We report the existence of 51,197 ERV-derived promoter sequences that initiate transcription within the human genome, including 1,743 cases where transcription is initiated from ERV sequences that are located in gene proximal promoter or 5' untranslated regions (UTRs). 114 of the ERV-derived transcription start sites can be demonstrated to drive transcription of 97 human genes, producing chimeric transcripts that are initiated within ERV long terminal repeat (LTR) sequences and read-through into known gene sequences. ERV promoters drive tissue-specific and lineage-specific patterns of gene expression and contribute to expression divergence between paralogs. These data illustrate the potential of retroviral sequences to regulate human transcription on a large scale consistent with a substantial effect of ERVs on the function and evolution of the human genome.

Introduction

Approximately 5% of the human genome sequence is derived from retroviruses [1]. Retroviral genomic sequences are remnants of past infections that resulted in the

integration of provirus genomes into the DNA of germline cells [71, 72]. The abundance of these so-called endogenous retrovirus sequences (ERVs) testifies to the extent that human evolution has been shaped by successive waves of viral invasion [73].

One way that ERVs have affected the function and evolution of the human genome is by donating regulatory sequences that control the expression of nearby genes. The gene regulatory effects of ERVs were first uncovered in a number of anecdotal studies on specific genes [74]. For instance, the long terminal repeat (LTR) of a human ERV (HERV-E) was shown to serve as an enhancer element that confers parotid-specific expression on the amylase gene [75]. Later, more systematic computational analyses of the human genome sequence revealed that many human genes contained ERV-derived regulatory regions, suggesting an even greater contribution of retroviruses to human gene regulation [68, 70]. Continued efforts to characterize ERV-derived promoters have turned up several new cases in recent years [76-78]. Nevertheless, the full extent of the contribution of ERV sequences to the initiation of transcription in the human genome has yet to be appreciated.

Initiation of transcription by ERV promoters often results in the production of alternative transcripts that are both tissue-specific and lineage-specific. For instance, testis-specific expression of the human gene encoding the neuronal apoptosis inhibitory protein (NAIP) is driven by an LTR promoter sequence, whereas a distinct LTR promoter in rodents confers constitutive expression of the orthologous gene [78]. An ERV LTR sequence also serves as an alternative promoter that drives expression of the beta1,3-galactosyltransferase 5 gene specifically in colorectal tissue [76].

The lineage-specific regulatory effects of ERV promoters can be attributed the fact that ERV sequences result from past germline infections, many of which occurred relatively recently along specific evolutionary lineages. In fact, most of the ERV sequences in the human genome are primate-specific [73], while most human genes are far more ancient and share orthologs with distantly related species [1]. This means that regulatory effects exerted by ERV promoters will often lead to expression differences between primate and non-primate orthologs or between deeper evolutionary lineages for more ancient ERVs. In other words, ERV promoters are likely to drive evolutionary changes in gene expression, long thought to be an important determinant of species divergence [79].

The application of novel high-throughput techniques for the analysis of gene expression has revolutionized the study of the human transcriptome and revealed far more regulatory complexity than previously imagined. Two techniques in particular, Cap analysis of gene expression (CAGE) and Paired-end ditag (PET) sequencing, enable the precise genome mapping of many thousands of promoter sequences that initiate transcription. CAGE is a technique that allows for the characterization of short sequence tags from the 5'-most ends of full-length cDNAs [30]. Accordingly, mapping CAGE tags to the human genome unambiguously identifies transcription start sites (TSS) and their corresponding promoters. PET sequencing involves the determination of sequences for tags from both the 5' and 3' ends of full-length cDNAs [80]. Thus, when PETs are mapped to the genome, paired transcriptional initiation and termination sites are identified along with the intervening genomic sequences that are transcribed as pre-mRNAs. We used human CAGE and PET data to more thoroughly evaluate the contribution of ERVs to the initiation of transcription in the human genome.

Methods

CAGE tags (n=1,551,672) were downloaded from the Japanese National Institute of Genetics website (http://genomenetwork.nig.ac.jp/public/download/cage_Database_e.html) and mapped to the human genome as previously described [30]. The human genome locations of PETs (n=669,840) were taken from the UCSC Genome Browser [55] annotations (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=100351785&c=chr9&g=wgEncodeGisRnaPet). The PETs were generated from several cell lines: log phase of MCF7 cells (113,858), MCF7 cells treated with estrogen (4,911), HCT116 cells treated with 5-fluorocil (124,770) and log phase of embryonic stem cell hES3 (426,301). Overlapping CAGE tags and overlapping PETs were clustered to identify individual TSS on the human genome. The UCSC Table Browser [81] and the program Galaxy [64] were used to compare the locations of CAGE tags and PETs to the locations of human ERVs annotated with the RepeatMasker program [5]. Only ERVs *sensu strictu*, as opposed to more ancient mammalian apparent LTR-retrotransposons (MaLR), were analyzed here. The National Center for Biotechnology (NCBI) Refseq [82] gene model annotations were used to evaluate the production of chimeric transcripts that are initiated by ERVs and read-through into human genes. Transcriptional units (TUs) are defined as genomic regions spanning the 5' to the 3' ends of individual Refseq gene models. TUs and 1 kilobase (kb) flanking regions upstream and downstream of TUs were evaluated for the presence of ERV-derived promoters. A series of custom Perl scripts were used to post-process the genome mapping data and to produce browser extendable data (BED) mapping tracks for further

analysis with the UCSC Genome Browser. All scripts and mapping data are available upon request.

The genomic presence/absence of ERV insertions across species was evaluated using whole genome sequence alignments of complete mammalian sequences built with the Multiz tool [83]. Human genome sequence conservation levels are based on the phastCons tool [84]. The species distribution of human gene orthologs was assessed using BLASTP [85] results from the NCBI Blink utility along with homology annotations from the GeneCards webserver [86]. Gene expression analysis was based on the Novartis Gene Expression Atlas version 2 (GNF2) [87].

Detailed information on all methods including PET and CAGE analysis along with gene expression and Gene Ontology analyses can be found in the Supplementary Information.

Results and Discussion

A total of 49,814 mapped CAGE tag clusters, each corresponding to an individual TSS, were found to map to the ERV LTR sequences (Table 3.1). The high number of ERV-derived TSS in the human genome identified with CAGE tag mapping underscores the potential of retroviral promoters to drive transcription. However, it is not possible to directly assess whether retroviral promoters identified using CAGE tag mapping actually drive the expression of known human genes. In fact, most of the ERV promoters identified with CAGE map to intergenic regions. This intergenic ERV promoter activity is likely to be a relic of the ERVs' ability to drive transcription of their own genome sequences from LTR promoters and may not necessarily be related to the transcription of human genes. Nevertheless, the presence of widespread ERV promoter activity in the

human genome demonstrates that ERV sequences can maintain the ability to promote transcription for millions of years after their initial insertion into the genome.

Table 3.1. Numbers of ERV-derived TSS in the human genome

Data source	Total TSS[a]	Gene-associated TSS[b]
CAGE	49,814	9,292
PET	1,513	114

[a] Total number of tag clusters representing individual ERV-derived TSS.
[b] For CAGE data, ERV-TSS that map within 1kb upstream or downstream of Refseq gene annotated 5' UTR sequences. For PET data, ERV-ditag sequence clusters that start within 1kb of Refseq gene annotated 5' UTR sequences, or within 5' UTRs, and end within Refseq gene TUs, 3'UTRs or 1kb downstream of 3' UTRs.

In addition to the intergenic ERV promoters, there are 9,292 CAGE identified ERV promoters that initiate transcription within 1kb upstream or downstream of the previously characterized TSS of known human genes (Table 3.2). PET sequence mapping data were also used to search for transcripts that are initiated from ERV promoters, and there are 1,513 cases of PET identified ERV promoters in the human genome (Table 3.1). Because PET sequence tags include both the 5' and 3' ends of full-length transcripts, they can be used to identify transcripts that are initiated within ERV sequences and read-through into human gene regions. These cases correspond to chimeric transcripts, composed partially of both ERV and human gene sequences, and demonstrate ERV promoted expression of human genes. This approach identified 114 distinct retroviral TSS that promote transcription of human genes (Table 3.2 and Table A.1). 21 of these retroviral promoters have co-located ESTs, which independently support their ability to initiate transcription. These retroviral TSS correspond to 124

Refseq transcripts over 97 distinct gene loci. The positions of TSS for ERV-derived human gene promoters were analyzed to evaluate whether ERVs provide canonical promoters or promote alternative transcripts. While there are a number of ERV TSS that map to 5' UTRs (Table A.2 and Figure A.1), and are thus taken to promote transcription at (or near) previously characterized TSS, the majority of ERV promoters promote alternative transcription of human genes from upstream regions or from within the TU (Table A.2). This further underscores the fact that ERVs promote alternative transcription of human genes. The ability of ERVs to promote alternative transcripts of human genes is illustrated (Figure 3.1) by the case of an alternative promoter of the glutathione-S-transferase omega 1 encoding gene (*GSTO1* Refseq accession NM_004832) found on chromosome 10q25.1. The GSTO1 protein is a member of the theta class glutathione S-transferase-like family, and it has been shown to act as a stress response protein through cellular redox homeostasis [88]. *GSTO1* nucleotide polymorphisms have been implicated in a number of cerebrovascular diseases including Alzheimer disease, Parkinson disease, vascular dementia and stroke [89, 90].

Table 3.2. Numbers of ERV-human gene associated or chimeric transcripts.

CAGE[a]			
Total	Upstream	5' UTR	TU
9,292	193	1,550	7,549
PET[b]			
PET 3' ends	PET 5' ends		
	Upstream	5' UTR	TU
TU	5	6	34
3' UTR	12	13	21
Downstream	4	8	11

[a] Counts for ERV-derived CAGE sequence tag clusters that map within human Refseq gene 5' UTRs or 1kb upstream or downstream (i.e. within the TU) of the 5' UTR.
[b] Counts for ERV-derived PET sequence clusters associated with human genes are shown. ERV-PETs with 5' ends that are 1kb upstream of human Refseq gene 5' UTRs, within 5' UTRs or within TUs are shown in columns. ERV-PETs with 3' ends that are within TUs, in 3' UTRs or 1kb downstream of 3' UTRs are shown in rows
1,550 of these ERV CAGE tag clusters map to 5' UTRs, consistent with transcription from previously characterized promoters, but the majority (7,742) map just upstream of the 5' UTR, in the proximal promoter region, or downstream within genes' TUs. Therefore, these ERV-derived promoters are likely to be responsible for generating alternative transcripts of human genes.

There is an ERV LTR sequence from the MER4A subfamily of sequences less than 500bp upstream of the Refseq annotated 5' UTR of *GSTO1* (Figure 3.1A). There are 15 individual PET sequences, forming 3 distinct TSS clusters, that have 5' ends inside of the MER4A sequence and 3' ends in the 3' UTR of *GSTO1* (Figure 3.1A and 3.1B). All of the MER4A PET sequences were derived from only one of the four PET libraries (χ^2=8.6 P=0.04), log phase of embryonic stem cell hES3, indicating that this promoter is tissue- or condition-specific. In addition to the PET sequence based evidence, there are a number of spliced ESTs that also indicate the MER4A sequence as an alternative promoter for *GSTO1* (Figure 3.1B).

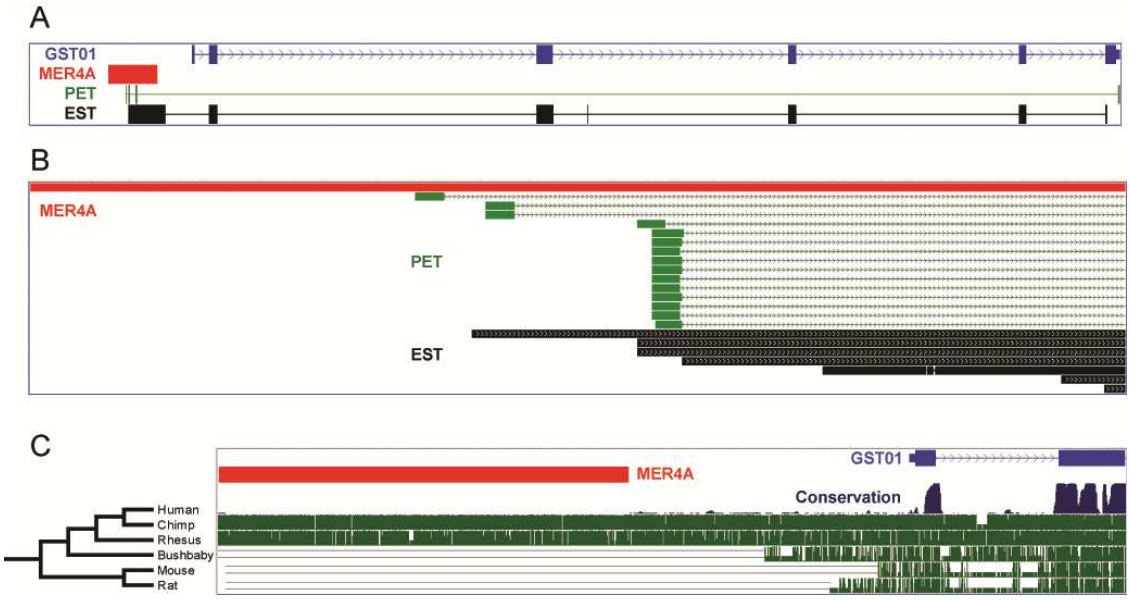


Figure 3.1. MER4A alternative promoter of the GSTO1 gene. A) TheMER4A (red) ERV sequence is located in the proximal promoter region <500bp from the GSTO1 5' UTR. The locations of PET sequences (green) and spliced ESTs (black) are shown. B) The MER4A (red) sequence region is enlarged and the individual PET sequences (green) and spliced ESTs (black) that support the existence of this promoter are shown. C) Evolutionary conservation of MER4A versus GSTO1. MER4A is only found in chimp and rhesus and no other mammals (green bars), i.e. it is not conserved, whereas the adjacent GSTO1 exons are conserved across mammalian species (green and blue bars).

Inspection of multiple sequence alignments of complete mammalian genomes reveals that the *GSTO1* adjacent MER4A insertion is present in the human, chimp (*Pan troglodytes*) and rhesus (*Macaca mulatta*) genome sequences but absent in the bushbaby (*Otolemur garnetti*), mouse (*Mus musculus*), rat (*Rattus norvegicus*) and all other placental mammal sequences (Figure 3.1C). In other words, that particular MER4A sequence inserted after the primate radiation began, and it is specific to the Haplorrhini suborder, which includes both new world and old world monkeys. On the other hand, the adjacent exonic sequences of *GSTO1* show marked conservation compared to MER4A (Figure 3.1C). Comparative sequence analysis with BLASTP indicates that *GSTO1* is far

more ancient than the MER4A insertion, having well conserved orthologs among mammals and other vertebrates along with *Drosophila melanogaster*, *Caenorhabditis elegans* and a number of other more distantly related species.

The comparative sequence analysis suggests the possibility that the MER4A insertion may confer lineage-specific expression pattern on *GSTO1*. Furthermore, there are two human paralogs of *GSTO1*, *GSTO2* and the pseudogene *GSTO3P1*, neither or which has the upstream MER4A insertion. So the specific regulatory effects of the ERV may not only be tissue- and lineage-specific but could also be involved in driving functional differentiation of paralogs via expression differences.

In order to test for potential diversifying regulatory effects of the MER4A insertion on *GSTO1*, we compared tissue-specific expression patterns between human and mouse *GSTO1* and *GSTO2* orthologs as well as between human *GSTO1-GSTO2* paralogs using microarray data from the Novartis Gene Expression Atlas version 2 (GNF2) [87]. The human-mouse *GSTO1* orthologous pair has a low (r=-0.06), and not significantly different from 0 (P=0.77), correlation of expression levels across tissues as does the human *GSTO1-GSTO2* paralogous pair (r=0.006 P=0.98) (Figure A.2). On the other hand, the human and mouse *GSTO2* orthologous genes, which lack the alternative MER4A promoter, have significantly correlated expression patterns (r=0.76 P=2.2e-6). These patterns of expression divergence and conservation are consistent with variation in expression introduced by the MER4A-TSS. In all, 37 out of 40 evaluated cases of human genes with ERV-TSS have expression patterns that are not significantly correlated with their mouse orthologs that lack the upstream ERV (Figure A.3).

We further evaluated the potential regulatory effects of human ERV-derived promoters by comparing the expression patterns of all human genes with ERV-promoters versus genes without ERV promoters using the GNF2 data. Human genes with ERV-TSS have greater tissue-specificity than genes lacking ERV promoters, consistent with a diversifying regulatory effect of ERV-TSS (Table A.2). In particular, ERV-TSS containing genes have anomalously high levels of expression, on average, in brain and testis (Figures A.4 and A.5). A similar pattern of significantly elevated expression in brain and testis was found for ERV CAGE tags (Figure A.6). Consistent with the brain-specific expression pattern of ERV-TSS genes, Gene Ontology (GO) functional analysis indicated that these genes are enriched for metabolic and signaling processes active in the brain (Table A.3 and Figure A.7).

Our analysis revealed that retroviral sequences in the human genome encode tens-of-thousands of active promoters; transcribed ERV sequences correspond to 1.16% of the human genome sequence and PET tags that capture transcripts initiated from ERVs cover 22.4% of the genome. These data suggest that ERVs may regulate human transcription on a large scale. However, it is a formal possibility that many of the ERV derived promoters identified here represent leaky transcription, *i.e.* noise, which is not functionally significant. Definitive proof of biological activity for individual ERV-TSS may have to await experimental confirmation via knock-out data or promoter swapping. However, it will soon be possible to validate ERV-TSS on a genome-scale owing to the accumulation of high-throughput data from tiling array experiments based on ChIP-chip and/or chromatin structure assays. Such data, which are being generated by the ENCODE Project Consortium (2007), measure the distributions of regulatory signatures

across genomic sequence. The presence and density of regulatory signals, such as transcription factor binding sites and open or specifically modified chromatin, have been shown to discriminate between biologically active and artifactual TSS and thus could be used to validate ERV-TSS.

Our analysis uncovered more than 100 cases of novel ERV-derived promoters that initiate chimeric ERV-human gene transcripts and several thousand more that are likely to do so. ERV-derived promoters are characterized by their ability to promote alternative transcripts that are expressed in a way that is tissue-specific, lineage-specific and distinct from related paralogous genes. These data underscore the extent to which retrovirus activity has shaped the human transcriptome.

CHAPTER 4

IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES DERIVED FROM TRANSPOSABLE ELEMENT SEQUENCES USING CHIP-SEQ

Abstract

Transposable elements (TEs) form a substantial fraction of the non-coding DNA of many eukaryotic genomes. There are numerous examples of TEs being exapted for regulatory function by the host, many of which were identified through their high conservation. However, given that TEs are often the youngest part of a genome and typically exhibit a high turnover, conservation based methods will fail to identify lineage- or species-specific exaptations. ChIP-seq has become a very popular and effective method for identifying *in vivo* DNA-protein interactions, such as those seen at transcription factor binding sites (TFBS), and has been used to show that there are a large number of TE-derived TFBS. Many of these TE-derived TFBS show poor conservation and would go unnoticed using conservation screens. Here, we describe a simple pipeline method for using data generated through ChIP-seq to identify TE-derived TFBS.

Introduction

Transposable elements (TEs) are segments of DNA that possess the ability to 'transpose,' meaning that they can move themselves to distant locations of the host genome and replicate when they do so. TEs are present in all domains of life, and abundant in the genomes of many sequenced eukaryotes accounting for a large portion of non-coding

DNA and the genomes as a whole (nearly 50%, ~1.4Gb of the human genome) [1]. Broadly speaking, there are two types of TEs. Type I TEs, or retroelements, transpose by a copy and paste mechanism via an RNA intermediate, generating a new insertion. Type II TEs, or DNA transposons, move by a 'cut and paste' mechanism where the actual insertion is moved [91]. Most TEs harbor their own promoters and regulatory sequences, and many active elements encode genes for their own transposition. Active elements are a small minority, however, and most TE insertions are unable to transpose.

Exaptation of Transposable Elements

TEs exist solely to continue their own existence; they do not, simply by their replication, contribute anything to the host [92, 93]. It is likely that many, if not the large majority of TE insertions, have little or no functional role for the host and are effectively under neutral or nearly neutral selection. However, given the very large number of TE insertions in eukaryotic genomes and the opportunistic nature of evolution, it is only reasonable to expect that some would be ‘exapted’ [94] over time to take on a functional role that benefits the host, a process that could have a wide variety of results [95, 96]. A key factor in TE exaptation events is their ability to promote their own transcription; without this ability, they could not replicate themselves. Given this ability, it stands to reason that TEs could be exapted to provide alternative promoters for host genes; this has been seen a number of times [97, 98]. Of most importance to this chapter, however, is the ability of TEs to provide new TFBS to the host. If there existed an active TE that contained a TFBS, then each new insertion that the TE generated would also contain the TFBS. If the TE were highly active, it could quickly spread the TFBS around the genome. Even if the TE simply had a sequence that was only close to the TFBS, it could

still spread this 'progenitor sequence' around the genome. Over time, point mutations in individual insertions could alter the progenitor sequence so that it would now be bound by the TF [99]. Either way, the TE could spread the TFBS around the genome over timer and create a network of TFBS, and in doing so alter the expression patterns of host genes. For example, it was recently shown that a large number of human c-myc binding sites are located in TE insertions, possibly creating a sub-network for c-myc control [100]. For a comprehensive review of TE-derived regulatory networks, see [2].

Transposable Elements Evolve Rapidly

Transposable elements are generally the most rapidly evolving part of a genome; so long as their insertions are not too deleterious to the host, TEs can quickly increase in copy number and then are generally free to accumulate point mutations. The rapid activity of TEs relative to the host genome means that lineage-specific insertions can be accumulated in a very short time frame. In the 6 MY since the human-chimpanzee divergence, for example, there have been several thousand new TE insertions in each genome . There also appears to be very little selective pressure on the deletion of most insertions, which can result in their chance deletion from one lineage, while they are retained in others. Between human and mouse, there is generally very little conservation of non-coding regions in the genome, including TEs. Many insertions that appear to predate the human-mouse divergence are present in one genome, but have been lost in the other (Figure 4.1.) [12]. The rapid insertion of TEs combined with their rapid loss means that two lineages can develop distinct TE complements in a relatively short time after divergence. Given that two lineages can have very different TE complements, it could be possible for a large number of lineage or even species-specific exaptation events (Figure

4.1). If the exaptation events were the creation of new TFBS or promoters, then the spread of TEs could create species-specific patterns of gene expression [69, 101].

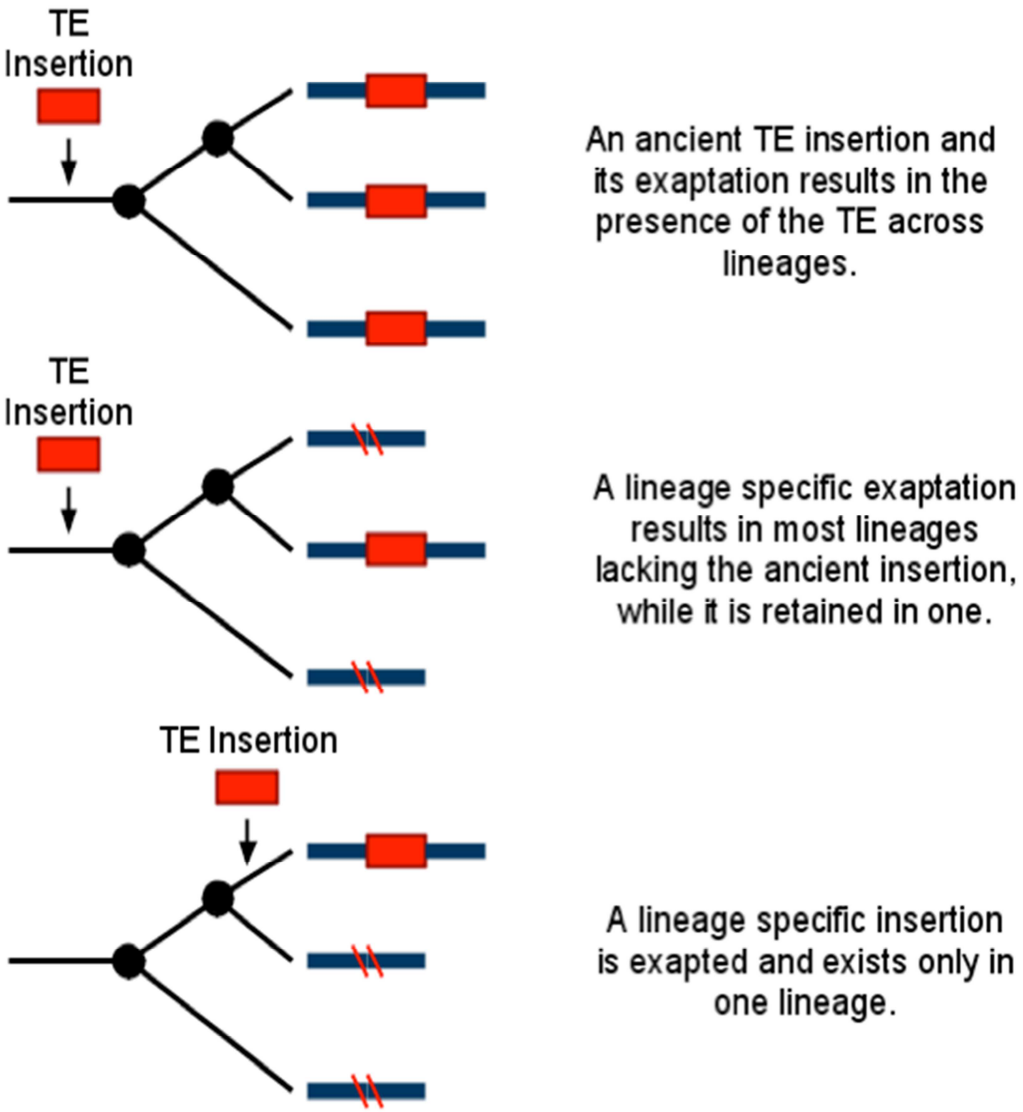


Figure. 4.1. Evolutionary scenarios related to TE exaptation events. A) An ancient insertion is exapted and the resulting regulatory sequences are shared across multiple derived evolutionary lineages. B) An ancient insertion is exapted but only selectively conserved in some of the derived evolutionary lineages. This could result in regulatory divergence between lineages. C) A recent lineage-specific insertion is exapted resulting in regulatory differences between lineages. TEs are particularly prone to this scenario given how dynamic and rapidly evolving they are.

Detection of Functional TE-derived Non-coding sequences

There are three widely used methods to find TFBS in genomes. It should be noted that these approaches are not mutually exclusive; indeed, the methods are often combined to more rigorously predict and locate TFBS. The first approach, phylogenetic footprinting, [102] can be done solely computationally via comparative sequence analysis. A phylogenetic screen attempts to find regions of different genomes that have been conserved over time, and in the case of TFBS, looking for conserved non-coding elements (CNEs). Screens looking for conserved non-coding elements (CNEs) represent a very successful technique for identifying the oldest, and due to their conservation most likely to be essential, non-coding parts of the genome. Shortly after the sequencing of the human and mouse genomes, it was shown that a larger than expected number of mouse MIR and L2 elements had human orthologs [12]. Subsequently, several thousand insertions or insertion fragments near human genes were shown to be under purifying selection, suggesting their exaptation and possible involvement in transcriptional control [103]. In recent years, a number of insertions have been shown to be enhancers for human and vertebrate genes, many identified with phylogenetic screens. An insertion from the CORE-SINE family was shown to be conserved across the mammalian lineage, and to be an enhancer of the POMC gene in mice [104]. The amniote SINE 1, AmnSINE1, family of TEs is a very old family that spread early in the amniote lineage. However, a number of conserved AmnSINE1 insertions exist in the human genome, two of which were shown to be enhancers involved in brain development [22, 23, 105]. A mammalian interspersed repeat (MIR) was shown to have enhancer 'boosting' activity, in that its presence greatly increased the action of a nearby enhancer, while the MIR could

not on its own be an enhancer [106]. The problem with an approach based on conservation is that, while it will find many important regions, the screen will miss other regions that are also important, but also lineage-specific. Lineage-specific TFBS, such as those that could be provided by lineage specific TE insertions, could generate lineage-specific expression, and would this be missed by CNE screens [101]. Another case in which older elements may be overlooked in CNE screens is one in where an old insertion has been lost, as many are, in several lineages, but exapted in one (Figure 4.1). Such an insertion may well play some role in the lineage that kept it, but it will be completely missed in CNE screens. CNE screens will not only miss new TE exaptations, but also other non-coding functional elements. It has been shown previously that sequences with low conservation can play important functional roles, such as rapidly evolving, long non-coding RNAs [107].

The second of the three methods to identify TFBS is also computational and involves scanning a genome for the sequence motif that the TF in question recognizes. REST, the RE1 Silencing Transcription factor, is known to repress neuronal genes in non-neuronal cells. Using experimentally identified REST binding sites, which contain the RE1 motif, Johnson *et al***.** [108] created a Position-Specific Scoring Matrix, PSSM, for the motif, and used it to screen for possible REST binding sites in the human genome. Johnson *et al*. were able to show that there are a number of TE-derived REST binding sites that had the ability to bind REST *in vitro*, suggesting that TEs have helped to spread the REST network. When a PSSM is used to search for new TFBS in a genome, false positives are controlled by shuffling the sequence in the PSSM, re-scanning the genome with the shuffled sequence and comparing the number of sites

identified with the original PSSM to those found with the shuffled PSSM [53]. This approach will not work, however, for TFs that recognize motifs smaller than the RE1 motif as there will likely be many false positives. In addition, the presence of a TFBS sequence motif does not guarantee that the sequence that bears it is actually bound by its corresponding TF, while sequences that lack similarity to the motif may in fact be bound by that factor. These challenges to the sequence-based computational approach necessitate an approach to identifying TFBS on a genome wide scale that does not depend on the sequence of the TFBS, only the binding of the TF to the region.

The third major approach to finding TFBS is identifying *in vivo* protein-DNA interactions via chromatin immunoprecipitation (ChIP) followed by microarray analysis (ChIP-chip) or sequencing of the captured DNA. Of the three approaches, this one offers the greatest sensitivity and potential specificity. ChIP is able to find genomic DNA that is bound by a transcription factor, not just those regions that are conserved or for which there exists a well-defined TFBS motif. ChIP is also distinguished from the other approaches in the sense that it identifies sequences that are experimentally characterized to be bound by transcription factors, *i.e.*, not just computational predictions. Genome-wide ChIP assays, such as ChIP-PET or ChIP-chip have been used successfully in the past; however, a newer and relatively inexpensive method, ChIP-seq has quickly become the dominant method of experimentally identifying TFBS, and it is on ChIP-seq that we focus the rest of our discussion. The ChIP-seq method combines ChIP with massively parallel sequencing of the bound DNA [27]. The sequencing is usually carried out on one of the currently available short-read sequencers: Illumina Genome Analyzer, ABI SOLiD, or Helicos HeliScope. ChIP-seq has a number of advantages over ChIP-chip and

ChIP-PET. There is no cross-hybridization, as can occur in ChIP-chip, and the ChIP-seq signal is a digital count of reads mapping to the TFBS, rather than a fluorescence signal. ChIP-seq is also far less costly than ChIP-PET, which typically relied on capillary sequencing. Using several ChIP-based data sets, including one derived with ChIP-seq, Bourque *et al.* [109] identified a large number of TE-derived TFBS. The majority of TFBS they observed were not well conserved, with many being lineage-specific. This strongly suggests that expansion of TEs within a genome can lead to the concurrent expansion of transcription regulatory networks. Below, we provide a specific example detailing how analysis of ChIP-seq data can be used to identify TE-derived TFBS.

Software

All the software we describe and recommend here is publicly available.

Bowtie [110] http://bowtie-bio.sourceforge.net/

MuMRescueLite [111] http://genome.gsc.riken.jp/osc/english/dataresource/

UCSC Genome Browser [112] http://genome.ucsc.edu

UCSC Table Browser [81] http://genome.ucsc.edu

Methods

This section describes our choice of tools for the identification of TFBS derived from TE insertions using ChIP-seq data, and we show how these tools can be assembled into an analytical pipeline. The tools presented were chosen for their speed, utility for analysis of TE-derived TFBS, ease of use and good documentation. To illuminate the use of these tools, we first provide an overview of our analytical pipeline for the detection of TE-derived TFBS (Figure. 4.2), and then we give a specific example of how ChIP-seq data can be analyzed to yield genome-wide set of TE-derived TFBS.

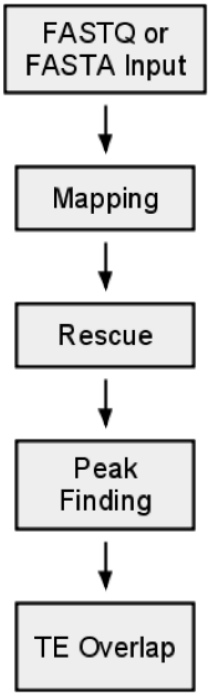


Figure. 4.2. Schematic of the analytical pipeline presented here for finding TE-derived TFBS with ChIP-seq. Each individual step is described in detail in the text along with important caveats, which are listed in 'Notes' section.

Mapping

The first step in finding TE-derived TFBS is to map reads generated by ChIP-seq back to the genome used. Massively parallel sequencers generate millions of reads in run of a ChIP-seq experiment. Mapping these reads in a genome as large as the human or mouse genomes with traditional techniques like BLAST [85] or BLAT [113] quickly becomes computationally overly expensive. Fortunately, a number of programs have been developed explicitly for the mapping of short-read data. The fastest of these are those that employ the Burroughs-Wheeler transform ***[114]*** to build a very dense index of the genome, then map reads using the index. We recommend Bowtie for general

mapping because of its speed and useful options (see Note 1). Bowtie is generally the fastest of these aligners, and it can utilize read quality information in the FASTQ format data generated from Illumina sequencing. However, it cannot currently use colorspace reads generated from SOLiD sequencing (see Note 2).

Read Rescue

Were genomes fully random sequences of the four bases, then almost any ChIP-seq read would be mappable to a unique region of the genome. However, due in large part to the vast number of TE insertions, this is not the case. There are numerous repeated sequences in eukaryotic genomes and sequence tags derived from these regions may not map unambiguously back to the genome – *i.e.*, they may map to multiple genomic regions with equal probability. The problem of such multiple-mapping ChIP-seq reads arises in part due to their short length. ChIP-seq reads must necessarily be short in order to provide good resolution protein binding locations in the genome; a 500bp read from ChIP-seq would be easy to unequivocal map to the genome, but would give very little information about the exact location of the DNA-protein interaction. A shorter read, on the order of <50bp, as most ChIP-seq data sets contain, gives good resolution regarding the location of the DNA-binding, but will have a much greater probability of mapping to multiple locations in the genome. If a TE insertion provides a TFBS, the insertion is very young, and there are many similar TEs in the genome, then it may not be possible to map the ChIP-seq reads from that insertion. For slightly older elements, there will be far fewer possible places to map the reads. Many studies have simply discarded multi-mapping reads for both simplicity of analysis, and a desire to be conservative in their findings. However, this becomes an obvious problem when

studying TEs, as this will result in the loss of many of the reads coming from TE insertions. To appropriately analyze ChIP-seq data in regards to TEs, some ‘rescue’ method must be used to resolve reads the map to multiple locations.

Different Methods of Rescue

There are currently several different schools of thought regarding 'rescuing' reads that map to multiple genomic locations. MAQ [115] is a very commonly used mapping utility for short read data. When it encounters reads that map to multiple locations with equal probability, it randomly chooses one of the locations to map the tag. This poses problems for TE-derived sequences, as it will dilute the signal from legitimate TFBS, potentially resulting in both false positives and false negatives. This method also ignores information on the local context of potential map positions given by uniquely mapping reads. MuMRescueLite [111, 116] takes this information into account and assumes that multi-mapping reads are more likely to come from regions which already have more uniquely mapping reads, and probabilistically determines where a read most likely came from. We recommend that MuMRescueLite be used after the initial mapping to resolve multi-mapping reads.

Peak Calling

Quality mapping is critically important for downstream analysis, and once this has been achieved, the first step is often finding 'peaks' or, more generally speaking, regions, that have a density of mapped ChIP-seq reads significantly higher than the background (see Note 3). These peaks are the regions bound by the TF that is being looked at in the ChIP-assay, and should contain the TFBS. Methods for peak calling, and indeed the area itself, are still new, and while there is work to be done in the area, there are several

quality software choices available for identifying peaks in ChIP-seq data. PeakSeq [117] and SISSRs [118] are two widely used utilities, and in this review, we recommend SISSRs due to its good documentation.

Finding TE-derived TFBS

SISSRs attempts, and in general is highly successful at, finding the TFBS to within a few tens of base pairs based on the strand orientations of reads forming the peak, as well as the density of reads in the region. Ideally, the TFBS would always be at the point of highest read density. In reality, it is very often co-located with the highest density, or if not that then very near by and SISSRs is correct in its predictions the large majority of the time. What this means, practically, is that finding those regions identified by SISSRs that are contained within TEs will tell us which TFBS are TE-derived (see Note 4). This can be accomplished in a number of ways, the simplest being the creation of two BED-formatted custom tracks for the UCSC Genome Browser [112], one from the predicted TFBS and one from the TEs, and uploading them to the browser. Then, the table browser can be used to intersect the tracks (see Note 5). Below, we provide a specific step-by-step example of how this can be done using the software cited in **Section 2.** Software.

Example

Here we provide an example using ChIP-seq data for the CCCTC-binding factor (CTCF) from the human ENCODE (ENCyclopedia of DNA Elements) project [25]. CTCF is zinc finger binding protein with multiple regulatory functions including both transcriptional activation and repression as well as insulator and enhancer blocking activity [119]. The ChIP-seq data for CTCF are available at

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap/. For this example, we will be using the first repetition of CTCF and the control. The majority of the steps in this procedure are done from the command line in the Unix/Linux operating system environment.

Mapping

The program Bowtie requires an index for the genome that the user wishes to map the tags to. This is accomplished with the 'bowtie-build' utility. It takes as input a FASTA file that contains the genome in question, the human genome in our example:

```
$bowtie-build <human genome FASTA> <index name>
```

Building the index typically takes several hours depending on the machine, though once built there is no need to build it again for different samples. Bowtie takes as input a FASTQ file and the parameters to control the mapping (see Note 1), as well as the index to use for the mapping:

```
$bowtie –q –v 4 –k 10 –m 10 --best --strata <index name> <FASTQ> <bowtie output>
```

The mapping should be done for both the CTCF ChIP-seq set and the control set. Bowtie is capable of mapping several thousand reads per second, or far more, depending on how many cores it is allowed to use (see Note 1).

Multi-mapping Read Rescue

MuMRescueLite takes all of the information that the Bowtie output has, but the information needs to be rearranged to meet the requirements of MuMRescueLite:

```
$awk  '/./  {print  $1"\t"$7  +  1"\t"$3"\t"$2"\t"$4"\t"$4  +  length($5)"\t1"}3333
       ' < <bowtie output> > <MuM Input>
```

While the above command may appear daunting, it is simply using awk to rearrange the columns of the Bowtie output and put tabs between them. MuMRescueLite is invoked with a much simpler command:

```
$MuMRescueLite.py <MuM Input> <MuM Output> <Window Size>
```

Keeping the window size small will prevent distant reads from rescuing reads that do not really come from the location. We suggest keeping the window size under 100. MuMRescueLite produces output that is the same as the input, with an additional column that represents the calculated probability that the read in question is from that site. Using the desired probability cutoff for multi-mapping read, use awk to create a BED track from the MuMRescueLite output for analysis with SISSRs:

```
$awk ‘$8 > <cut off> {print $3”\t”$5”\t”$6”\t”$4}’ <  <MuM Output> >  <Mapping
BED>
```

The output should then be sorted by chromosome, then start, then stop:

```
$sort –k 1,1 –k 2n,2n –k 3n,3n –o <Mapping BED> <Mapping BED>
```

As with the mapping, the rescue should be done for both sets.

Peak Calling

SISSRs takes as input the two BED files created in the previous step, and creates another file with peak calls:

```
$sissrs.pl –i <CTCF File> -b <Control File> -o <Output File>
```

Use of the `-i` option to specifies the ChIP set as the input, and the -b option to specify the control set as the background. The `-o` option tells SISSRs where to write the output. Formatting the output into a BED file will allow overlap of the identified TFBS with TEs in the UCSC genome browser:

```
$awk '/^chr/ {print $1,$2,$3}' < <Output File> > <TFBS BED>
```

Identification of TE-derived TFBS

The final step is to upload the SISSRs-identified TFBS, BED-formatted track to the UCSC genome browser as a custom track. The name of the track should be changed so as not to be overwritten by later tracks. Once that is done, create another custom track that will contain only TEs using the table browser. This can be done by filtering the RepeatMasker track for only those repeats which have a 'repClass' of 'LINE', 'SINE', 'LTR', or 'DNA.' Intersecting the track of CTCF TFBS with this TE-only track will give those TFBS that reside in TE insertions. If everything has gone right, then there should be examples like that shown in Figure 4.3. Here, two distinct CTCF binding sites are shown for a solo long terminal repeat sequence from the endogenous retrovirus family K (ERVK). Although these particular binding sites were identified solely based on ChIP-seq data, they can also be seen to possess known CTCF binding site sequence motifs at

the bound genomic intervals. Thus, a computational survey of TE sequences that possess TFBS motifs may have turned up this example.

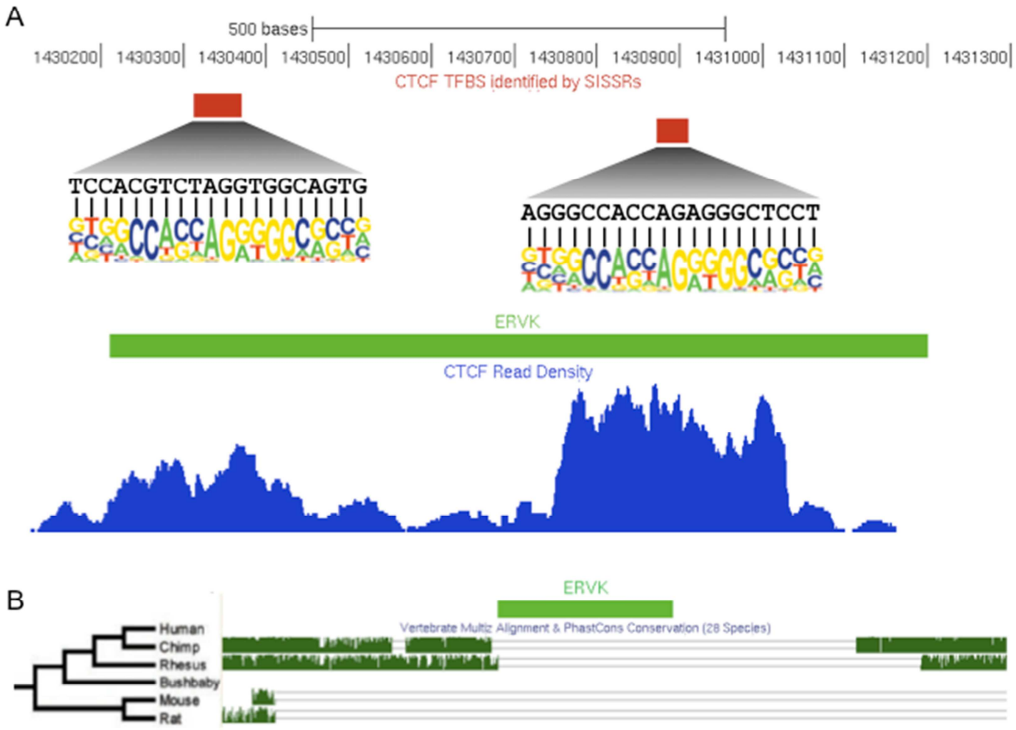


Figure 4.3. An example of two TE-derived CTCF binding sites found using ChIP-seq data. A) Two CTCF TFBS (red) identified by the SISSRs program are found within the long terminal repeat sequence of an endogenous retrovirus TE (ERVK in green). The ChIP-seq read density (blue) shows two peaks in the ERVK that correspond to the CTCF bound regions. Analysis of the bound regions with a CTCF position weight matrix (PWM) [120] using the program CLOVER [121] confirms the presence of two conserved CTCF binding site sequence motifs in the regions identified with the ChIP-seq data. The sequences of the binding sites are shown compared to the sequence logo representing position-specific variation in the CTCF PWM. B) Regions orthologous to the ERVK insertion site from completely sequenced mammalian genomes were compared using the vertebrate Multiz alignment. Sequence regions conserved between species are shown in green. Regions flanking the ERVK element are conserved in other mammalian genomes, but the insertion itself is human-specific.

Genome-wide there are 326 CTCF bound sites located within ERVK sequences, and ERVK elements show more than an order of magnitude greater likelihood to be bound by CTCF than members of other ERV families. The number of CTCF bound ERVK sequences suggests that these TE-derived TFBS may play some role in regulating human genes, and in fact many ERVs are located in close proximity to genes. For instance, the CTCF bound ERVK shown in Figure 4.3 is located in the 5' regulatory region ~6kb upstream of the ATAD3A gene.

ERV sequences in general, and members of the ERVK family in particular, are young lineage-specific elements that are poorly conserved across species. Phylogenetic analyses revealed that the ERVK family invaded the primate lineage subsequent to the diversification between New World and Old World monkeys [73]. Consistent with their recent evolutionary origin in the human genome, ERVK sequences have a mean PhyloP (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=147315896&c=chr1&g=phyloPCons28way) base-wise conservation score of 0.22, while the genome as a whole has a mean score of 0.47. Therefore, phylogenetic footprinting approaches, which identify regulatory sequences in non-coding DNA by virtue of their sequence conservation, would be exceedingly unlikely to turn up any cases of ERVK-derived TFBS. Indeed, comparison of the CTCF bound ERVK sequence shown in Figure 4.3 with orthologous mammalian genome sequence regions indicates that this particular element copy human-specific and missing in all other mammals. Such lineage-specific TE-derived regulatory sequences may be of particular interest in the sense that they could be responsible for driving regulatory divergence between species [69, 101].

Notes

1. Bowtie is currently the fastest short-read aligner available and our preference for mapping short-read data, such as that generated by ChIP-seq or RNA-seq. It has many of the same advantages of MAQ, such as taking quality information into account, but also has other features useful for looking at TE-derived sequences that MAQ currently lacks. Bowtie is also quite memory-efficient and it scales well with genome size. Bowtie can be run with the human genome on a computer with 4GB of RAM, though on such a computer nothing else should be started in the meantime, as when Bowtie is forced out of memory it tends not to recover. Bowtie has a large number of options for controlling mapping and output, which can be listed by executing bowtie with no arguments. The more important options are listed and explained here

`-v <integer>` This specifies that there can be only a certain number of mismatches in the whole length of the alignment between the read and genome, and not just in the seed as the default behavior allows. This is also important for resolving multi-mapping reads. We suggest setting it fairly high in case some bases in the read have wrong calls and low quality scores.

`-k <integer>` This option is critically important among those available. This option tells bowtie that it should report more than one mapping, as by default it reports only the first. At the current time, MAQ will not report more than one mapping. Currently, MAQ will use the quality scores to choose a location and assign the mapping a quality of 0. Output of multi-mapping reads and their possible location is essential the rescue and analysis of TE-derived sequences.

`--best` Giving this option will cause bowtie to report only those mappings which have the highest quality, and is recommended if you have the FASTQ data and not just the FASTA data of base calls. This can greatly reduce the number of multi-mapping reads.

`--strata` This option is used along with the --best option, and will cause bowtie to return only the highest quality mappings .

`-m <integer>` will eliminate reads that map more than m times. We suggest making it the same as k. This will remove reads that map to so many places in the genome that they could likely never be placed with confidence.

One major advantage of Bowtie is that it allows for the easy use of multiple cores, which every desktop shipped in the last ~3 years has. Speed will become increasingly important as the number of reads generated per run increases. On a dual-core machine, such as a machine with an Intel Core Duo, only 1 core is advisable. However, on a quad-core machine, it is generally advisable to use 2 or 3 cores. On an eight-core machine 6 cores are recommended. The number of cores (processors) is set with the -p option. In some unfortunate cases, FASTQ files from a ChIP-seq experiment are not available, and only the base calls are supplied. In this case, you would not supply the '-q' flag to indicate FASTQ format. It is in these cases that the rescue is especially important.

2. The ABI SOLiD sequencing platform does not produce base calls like the Illumina platform, but rather 'color' calls that represent transitions between two bases. Bowtie cannot currently map colorspace reads, and we suggest the SOCS program for this

purpose ***[122]***. Like Bowtie, it has generally low memory requirements and is also capable of using multiple cores when available.

3. Though many peaks from ChIP-seq data will be quite large and obvious, others may be closer to the background noise. Complicating this is that the background in ChIP-seq is non-random, and tends to form peaks of its own. Most peak-finding utilities will look for peaks with just the ChIP-seq data alone, but many also allow the use of both the ChIP-seq data and a control set. By comparing the control set and the experimental set, false positives that result from peaks not related to the ChIP can be removed.

4. While SISSRs and other peak finders do a very good job of finding the actual TFBS from ChIP-seq data, they may still be off on occasion. A more accurate way to find the exact TFBS is to scan the identified TFBS, along with their flanks, with a PSSM for the TFBS motif with a program such as MAST ***[123]***. This will give the exact location of the TFBS if it exists in the peak region.

5. In this Chapter, we suggest using the UCSC Genome Browser and table browser for the overlap of the identified TFBS and transposable elements. This is very simple to do, but requires loading BED-formatted tracks to the browser and (relatively) lots of manual work. 'Kent Source Tree' is a large series of utilities, many of which form the back end of the browser. One such utility, 'bedOverlap' will overlap two sets of tracks without having to upload them to the browser. Numerous other useful utilities include the 'bedItemOverlapCount' utility, that can produce custom 'wiggle' tracks for the UCSC

genome browser, which visualize the density of ChIP-seq reads, and hence protein binding intensity, along the genome. Compilation and installation of the Kent Source Tree is not always easy, but is recommended if possible.

Acknowledgements

I. King Jordan was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). I. King Jordan and Andrew B. Conley were supported by NIH HG000783 granted to Mark Borodovsky. The authors would like to thank Jianrong Wang for help with the CTCF analysis.

CHAPTER 5

EPIGENETIC REGULATION OF HUMAN CIS-NATURAL ANTISENSE TRANSCRIPTS

Abstract

Mammalian genomes encode numerous cis-natural antisense transcripts (cis-NATs). The extent to which these cis-NATs are actively regulated and ultimately functionally relevant, as opposed to transcriptional noise, remains a matter of debate. To address this issue, we analyzed the chromatin environment and RNA Pol II binding properties of human cis-NAT promoters genome-wide. Cap analysis of gene expression (CAGE) data were used to identify thousands of cis-NAT promoters, and profiles of nine histone modifications and RNA Pol II binding for these promoters in ENCODE cell types were analyzed using chromatin immunoprecipitation followed by sequencing (ChIP-seq) data. Active cis-NAT promoters are enriched with activating histone modifications and occupied by RNA Pol II, whereas weak cis-NAT promoters are depleted for both activating modifications and RNA Pol II. The enrichment levels of activating histone modifications and RNA Pol II binding show peaks centered around cis-NAT transcriptional start sites, and the levels of activating histone modifications at cis-NAT promoters are positively correlated with cis-NAT expression levels. Cis-NAT promoters also show highly tissue-specific patterns of expression. These results suggest that human cis-NATs are actively transcribed by RNA Pol II and that their expression is epigenetically regulated, pre-requisites for a functional potential for many of these non-coding RNAs.

Introduction

In recent years it has become evident that substantial portions of mammalian genomes are actively transcribed as non-coding RNA, including thousands of cis-natural antisense transcripts (cis-NATs) [33, 34, 42, 43, 124]. Cis-NATs are transcripts produced from within protein coding loci, but from the opposite strand, and are thus complementary to the sense mRNA transcript (Figure 5.1A). Cis-NATs may play important regulatory roles via transcriptional interference caused by collisions of RNA polymerase complexes moving in opposite directions across the same locus [34, 44] or through the formation of double stranded RNA leading to post-transcriptional silencing through RNA interference [125, 126]. However, the extent to which non-coding RNAs in general, and cis-NATs in particular, are biologically functional remains a matter of debate. Some studies have suggested that the majority of non-coding RNA transcripts are non-functional and simply represent transcriptional noise [127, 128], while others have found evidence in support of function for numerous non-coding RNAs [19, 129, 130].

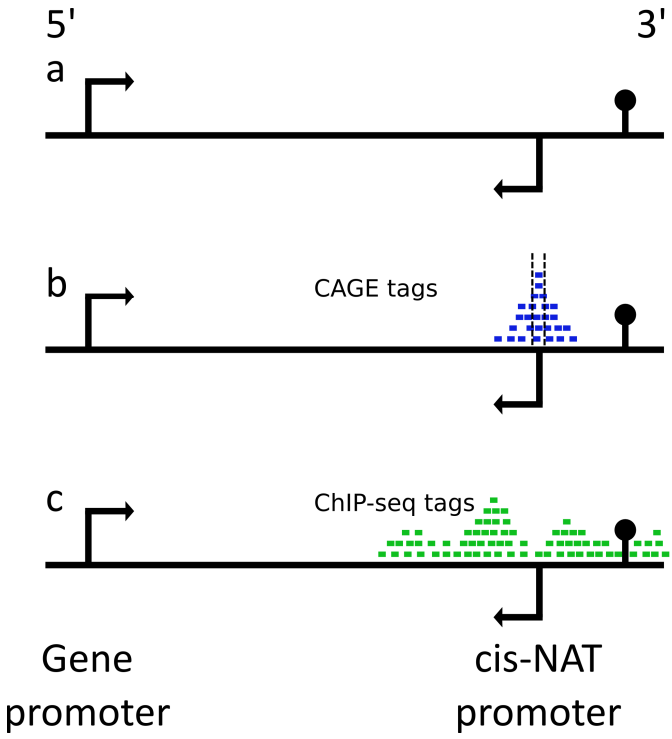


Figure 5.1. Delineation and analysis of cis-NAT promoters. Cis-NATs are initiated from within protein coding gene loci and transcribed in the opposite (antisense) direction. (a) Example of a protein coding gene locus with a genic promoter that drives transcription in the 5'-to-3' direction along with a cis-NAT promoter that initiates 3'-to-5' transcription within the locus. (b) cis-NAT transcription start sites (TSS) were defined using clusters of overlapping antisense CAGE tags. Specific cis-NAT TSS locations were taken as the base with the highest density of mapped CAGE tags within the cluster. (c) cis-NAT promoter sequences were taken as genomic regions immediately flanking cis-NAT TSS, and the chromatin environment of cis-NAT promoters was analyzed using ChIP-seq data for histone modification and RNA Pol II binding.

Previously, investigators have interrogated the functional potential of novel non-coding RNA transcripts by evaluating the chromatin environment in-and-around their promoters [19, 25, 131]. These studies were motivated by the fact that the promoters of well-characterized human genes have characteristic chromatin properties, including distinct protein binding and histone modification profiles, and these particular chromatin environments give indications as to the biological mechanisms, both genetic and

epigenetic, by which the genes are regulated [19, 25, 131]. For example, chromatin immunoprecipitation (ChIP-seq) studies have revealed that the promoters of actively transcribed genes are occupied by RNA Pol II and marked with a suite of specific histone tail modifications, such as acetylation of the lysine at position 9 of histone H3 (H3K9Ac) [3, 4, 132], whereas silent gene promoters are depleted for RNA Pol II and enriched for known repressive modifications such as trimethylation of lysine 27 of histone H3 (H3K27Me3). On the other hand, it has been shown that the promoters of many novel non-coding transcripts that have been characterized by high-throughput sequencing methods, but for which there is no additional supporting information, do not show enrichment for histone modifications or an active chromatin environment [19, 25, 131]. Thus, chromatin can be used to discriminate between the promoters of actively regulated genes versus putative TSS that probably represent transcriptional noise.

In this study, we evaluated the chromatin environment surrounding hundreds of thousands of human cis-NATs across six different ENCODE cell types for ten RNA isolation conditions. We sought to establish whether or not cis-NAT promoters show patterns of activity and chromatin modifications that are consistent with epigenetic regulation. We found that active cis-NAT promoters are enriched with active histone modifications and occupied by RNA Pol II, whereas silent cis-NAT promoters are depleted for both active modifications and RNA Pol II and enriched for the repressive modification H3K27Me3. These data provide evidence for the epigenetic regulation of numerous human cis-NATs, presumably a pre-requisite of their potential function as gene regulators.

Methods

CAGE data analysis

Human cis-NAT promoters were delineated using CAGE data from the ENCODE repository on the UCSC genome browser [28]. CAGE data from six cell types and across ten RNA isolation conditions were used for this study. The cell types are: GM12878, H1HESC, HepG2, HUVEC, K562 and NHEK. The RNA isolation conditions consist of polyadenylated and non-polyadenylated RNA fractions from whole cells, cytoplasm, nucleus, nucleolus and nucleoplasm. Altogether, a total of 16 different CAGE data sets were analyzed here (Table 5.1 and Table B.1). CAGE tags from each data set mapped to the reference sequence of the human genome (NCBI build 36.1; UCSC version hg18) [133] were clustered by their genomic locations to identify promoters. CAGE clusters with two or more co-located tags have previously been shown to represent validated transcription start sites (TSS) [32, 54]; accordingly, CAGE clusters containing two or more overlapping tags were used for the promoter analyses reported here. For each CAGE cluster, the actual TSS was characterized by finding the base with the highest density of mapped CAGE 5'-ends (Figure 5.1B). CAGE clusters that were anti-sense to a protein-coding locus from the UCSC known genes set were taken to be cis-NAT promoters as previously described [124] (Table B.2). To reduce contamination of the cis-NAT TSS by the possible degradation products of mRNAs, all CAGE clusters that overlapped an exon of the UCSC gene set were removed from the set of cis-NATs. CAGE clusters within 250bp of an annotated TSS of a protein coding loci were taken to be genic promoters and the TSS taken as the base with peak CAGE tag density. As a control, CAGE clusters that overlapped an exon of the UCSC gene set and were in the same orientation as the exon were kept for analysis.

ChIP-seq data analysis

Histone modification and RNA Pol II occupancy for cis-NAT promoters were evaluated using ChIP-seq data from the ENCODE repository on the UCSC genome browser [133]. Where available, FASTQ ChIP-seq data for the H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K9Me1 H3K27Me3, H3K27Ac, H3K36Me3 and H4K20Me1 modifications in the GM12878, H1HESC, HepG2, HUVEC, K562 and NHEK cell types, were taken from the ENCODE repository. A non-specific input ChIP-seq control data set was also analyzed for each of the ENCODE cell types. All ChIP-seq data were mapped to the May 2006 build of the human genome reference sequence (NCBI 36.1; UCSC hg18) using BowTie [110], keeping the best alignments with ties broken by quality scores. Any reads with more than 20 possible mappings were discarded. Remaining reads with multiple, high quality mappings were resolved using GibbsAM [134] (Table B.3). Tag counts for a given modification were normalized by dividing by the total number of mapped tags for that modification, then multiplying by ten million. ChIP-seq data were used to characterize the chromatin environment proximal to CAGE-characterized cis-NAT promoters (Figure 5.1C).

Association mining analysis

For each cell type we used only the CAGE data from the nucleus (GM12878, HEPG2, K562 and NHEK), cytosol (HUVEC) or whole cell (H1HESC) isolate to classify the activity of sense genic promoters in relation to the sum cis-NAT activity for the genic promoter, *i.e.* the sum of downstream cis-NAT promoter activity. For each cell type, genic promoters which had CAGE tags associated were ranked by their CAGE tag counts, and the top 25% were classified as ‘high activity’ in the cell type, while genic

promoters that had no CAGE data or were in the bottom 25% were classified as 'low activity' in the cell type. The same was done for the cumulative downstream cis-NAT activity of the genic promoters. This resulted in four possible classification combinations for cis-NAT and genic activity levels: 1) high cis-NAT & high gene, 2) high cis-NAT & low gene, 3) low cis-NAT & high gene, 4) low cis-NAT & low gene. We then used association mining to calculate the value of the *Interest* (*I*) parameter, as previously described [135], which is the ratio of the observed frequency of co-occurrence of any two classifications divided by their expected co-occurrence based on random association.

Statistical analysis

Student's *t*-tests were used to compare differences in the average number of normalized ChIP-seq tags +/-5kb of cis-NAT promoters for different cis-NAT activity levels (Figure 5.2). We used the statistical software R for calculating the Spearman's rank correlation coefficients for all correlation analyses (Figures B.3-B.4). The statistical significance of Spearman's rank correlation coefficients r was determined using the Student's *t* distribution with *d.f* .= n-2 with the formula $t = r\sqrt{(n-2)/(1-r^2)}$ [136].

Results and Discussion

Large-scale identification of cis-NAT TSS

CAGE (cap analysis of gene expression) is a method for characterizing the 5'-end of RNA transcripts; genomic mapping of CAGE sequence tags identifies transcription start sites (TSS) and promoters [30, 31]. CAGE combined with high-throughput sequencing can identify many thousands of TSS, while at the same time quantifying their promoter activity via the number of reads mapping to each TSS. CAGE data were

analyzed as described in the Methods to identify cis-NAT promoters in the human genome for the 16 different combinations of ENCODE cell type and RNA isolation conditions analyzed here. The number of cis-NAT promoters identified in this way ranges from 11,650 to 313,003 across the ENCODE cell types (Table 5.1 & Table B.2). We evaluated whether the large differences in cis-NAT promoters identified across cell types were due to differences in the numbers of CAGE tags per library or differences in sequencing quality across libraries. Library-specific read count values and read quality scores are not significantly correlated with the numbers of cis-NAT promoters identified across cell types, suggesting that the differences observed do not result from the CAGE data abundance or quality.

Table 5.1. Numbers of cis-NAT promoters identified by CAGE clusters in each cell line, sub-cellular location and poly-adenylation state.

Cell Line	Sub-cellular location	Poly-A-	Poly-A+	Total
GM12878	Cytosol	24,107	---	---
	Nucleoplasm	---	---	165,430
	Nucleus	62,704	---	---
H1HESC	Whole Cell	67,216	---	---
HEPG2	Cytosol	33,862	---	---
	Nucleoplasm	---	---	214,364
	Nucleus	265,896	---	---
HUVEC	Cytosol	25,309	---	---
K562	Cytosol	164,399	30,867	---
	Nucleoplasm	---	---	79,677
	Nucleolus	---	---	112,308
	Nucleus	313,003	148,461	---
NHEK	Cytosol	11,650	---	---
	Nucleus	178,016	---	---

Enrichment of chromatin modifications and RNA Pol II at cis-NAT promoters

To characterize the relationship between local chromatin modifications and cis-NAT promoter activity, we analyzed the number of ChIP-seq tags from each histone modification, and RNA Pol II, proximal (+/- 5kb) to each cis-NAT TSS. The analysis of cis-NAT chromatin modifications was conducted for 16 combinations of six ENCODE cell types over ten RNA isolation conditions. Here, we present an example of these results for one cell type and condition (NHEK cis-NATs characterized from nuclear non-polyadenylated RNA); results for all other cell types and conditions are detailed in the Supplement. Cis-NAT promoters were binned into 4 equal sized bins based on their promoter activity, from lowest to highest activity, as measured by CAGE tag counts. Histone modifications and RNA Pol II occupancy were then compared for cis-NAT promoters with different levels of activity. Cis-NAT promoters showed significant increases in ChIP-seq tag counts for the activating histone modifications H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac with increasing levels of cis-NAT promoter activity (Figure 5.2).

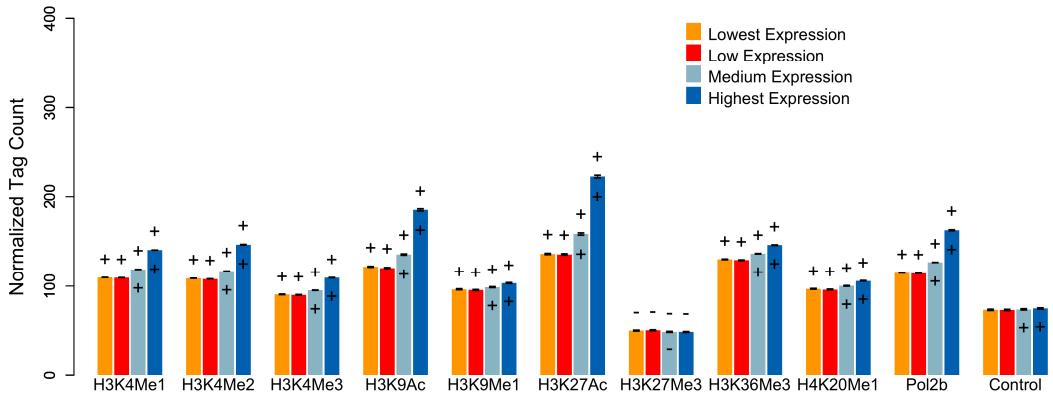


Figure 5.2. Enrichment of chromatin modifications and RNA Pol II at cis-NAT promoters. Cis-NAT promoters identified in the NHEK cell type were divided into 4 bins based on their activity (lowest to highest activity), and the normalized average numbers of ChIP-seq reads from each histone modification +/-5kb of the cis-NAT TSS were calculated for each bin. A '+' or '-' above a bar indicates that the number of ChIP-seq reads for that bin and modification is significantly higher or lower, respectively, than the control for that bin ($P < 0.001$). A '+' or '-' within the bar indicates that a bin is significantly enriched or depleted, respectively, for the histone modification compared to the next lowest activity bin ($P < 0.001$). Error bars shown are the standard error of the mean.

Furthermore, each of these modifications shows significantly greater average cis-NAT tag counts than seen for the ChIP-seq control (Figure 5.2). These histone modifications have previously been characterized as activating modifications by virtue of their association with the promoters of actively transcribed genes [3, 4, 132]. H3K27Me3, on the other hand, is known as a repressive modification that is associated with silent genes, and ChIP-seq tag counts for H3K27Me3 are lower than seen for the control in all cis-NAT promoter activity bins (Figure 5.2). Similar qualitative patterns are seen for H3K9Me1, H3K36Me3 and H4K20Me1, but the tag counts do not vary as much with cis-NAT promoter activity. This is may be due to the fact that these modifications are associated with transcribed regions, where the cis-NAT promoters are located, as opposed to promoter regions *per se* [3, 132]. In other words, chromatin signals of

promoter activity for these marks may be obscured by the fact that they are enriched within gene bodies where the cis-NATs are located. Overall, the patterns of enrichment seen for histone modifications at cis-NAT promoters suggest that the cis-NATs identified here are epigenetically modified in accordance with their relative expression levels and are thus likely to be specifically regulated, which is a pre-condition for their functional relevance, as opposed to non-specific artefacts such as RNA degradation products. For all activity levels, the level of Pol II binding is higher than seen for the non-specific input control, suggesting that regions near cis-NAT promoters are bound by Pol II. Qualitatively similar patterns of histone modification and Pol II occupancy across different cis-NAT promoter activity levels were seen for 14 out of the 15 remaining CAGE data sets analyzed here; the only exception was the NHEK cytosol CAGE data set (Figures B.1 and B.2).

To further evaluate whether histone modifications were correlated with cis-NAT promoter activity, cis-NAT promoters were divided into 200 bins based on activity as measured by CAGE tag counts. Cis-NAT TSS CAGE tag counts were then compared to ChIP-seq proximal promoter histone modification and RNA Pol II tag counts using the Spearman rank correlation (Figures B.3 and B.4). Cis-NAT promoter activity and histone modifications generally showed positive correlations for the activating H3K4 methylations and H3K9 and H3K27 acetylations and weaker, though still positive correlations for the H3K9Me1, H3K36Me3 and H4K20Me1 modifications. A weaker negative correlation was seen for the repressive H3K27Me3 modification. As would be expected for actively transcribed promoters, there was also a positive and significant

correlation between cis-NAT promoter activity and RNA Pol II presence and RNA-seq read density.

Histone modification, RNA Pol II occupancy and transcription near cis-NAT promoters

The enrichment of activating histone modifications and RNA Pol II occupancy near active cis-NAT promoters suggests that cis-NAT expression is epigenetically regulated; however, this enrichment could result from cis-NATs being located in open chromatin regions inside gene bodies, and not from the promoters being specifically modified to regulate their activity. To evaluate this possibility, we analyzed the distribution of histone modifications and RNA Pol II occupancy around cis-NAT TSS. If the enrichment of chromatin modifications observed for cis-NATs is due solely to their location in open chromatin, then we do not expect to see any variability in enrichment along chromosomal regions surrounding cis-NATs. On other hand, actively regulated cis-NATs would be expected to show modification peaks centered around the TSS as has been seen for the promoters of protein coding loci [3, 4, 25, 132].

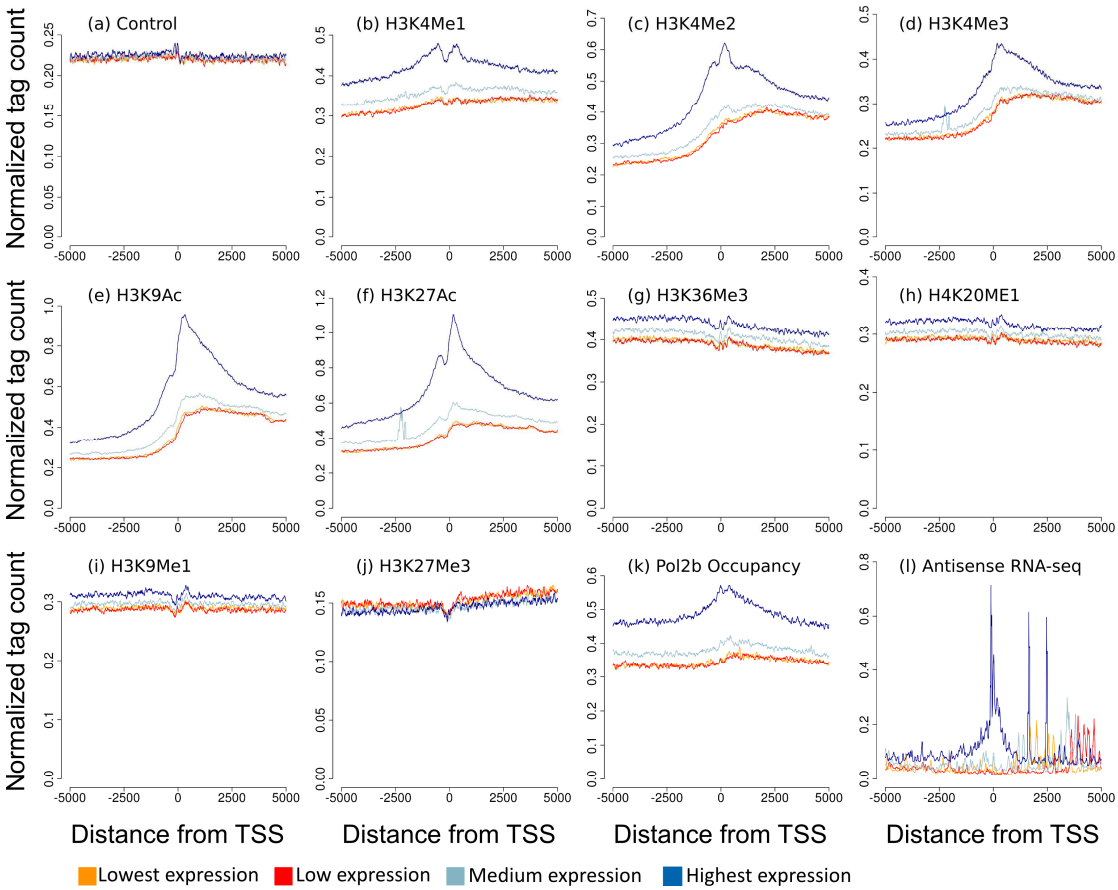


Figure 5.3. Chromatin modification and RNA Pol II environment around cis-NAT promoters. Cis-NAT promoters identified in the NHEK cell type were divided into 4 bins based on their activity (lowest to highest), and the normalized average numbers of ChIP-seq reads in 10 base-pair windows +/-5kb of the cis-NAT TSS (at position 0) were calculated for each bin.

Cis-NAT promoters were broken down by activity level, as described above, and the average numbers of ChIP-seq tags were calculated for 10 base-pair windows +/-5kb from cis-NAT TSS (Figure 5.3). Methylations of H3K4 (H3K4me1, H3K4me2 & H3K4me3) are known activating marks of promoters [3], and were all found to be enriched near cis-NAT promoters for the NHEK nuclear non-polyadenylated RNA data set (Figure 5.2). In further accordance with their epigenetic regulation, peaks of ChIP-seq read density from the H3K4Me1, H3K4Me2 and H3K4Me3 modifications were observed on either side of the cis-NAT TSS in this same data set, with more active promoters being more highly modified on average (Figure 5.3b,c,d). A notable dip can be seen near the cis-NAT TSS for these three modifications, suggesting nucleosome absence, similar to what has been seen at canonical TSS in $CD4^+$ T-cells [4]. Similar patterns were seen for the activating acetylations of H3K9 and H3K27 (Figure 5.3e,g). No discernable difference between bins was seen for the repressive mark H3K27Me3 (Figure 5.3j). The similarities seen for the genomic distributions of cis-NAT promoter modifications to those of protein-coding loci promoters [3, 4, 25, 132] provides evidence that cis-NAT expression is not simply transcription resulting from open chromatin, but is specifically regulated. The nucleosome absence seen even at the TSS with the lowest activity suggests that these TSS, which are identified by only a small number of CAGE tags, are *bona fide* TSS that have been epigenetically silenced by histone deacetylation. Pol II occupancy is seen at the TSS for all activity bins, with the higher activity bins showing a much higher occupancy, in accordance with the activity of the bins (Figure 5.3k). The H3K9Me1, H3K36Me3, and H4K20Me1 modifications show levels of

enrichment similar to the control with no observable enrichment on either side of the TSS (Figure 5.3g,h,i). This is likely to be due to the fact that these modifications are associated with actively transcribed regions, such as gene bodies, where the cis-NAT TSS in this study are located [3, 132]. RNA-seq data also peaks near the cis-NAT promoters, and increases with cis-NAT promoter activity (Figure 5.3l). Patterns of modification near cis-NAT TSS using CAGE and ChIP-seq data were qualitatively similar for 10 out of the 15 remaining CAGE data sets analyzed here; the HepG2 nucleus, K562 nucleoplasm, and both K562 nucleus CAGE sets have greatly distorted patterns of modification (Figures B.5 and B.6). Taken together, these data indicated that cis-NAT promoters show genomic distributions of histone modifications and RNA Pol II binding around TSS that are consistent with specific activation of transcription at the TSS as opposed to a simple accumulation of activating marks inside actively transcribed protein coding gene regions.

For comparison, the same chromatin enrichment analyses were done for CAGE clusters associated with genic promoters in the 6 ENCODE cell types. The patterns of local histone modifications for these promoters were largely qualitatively similar to those seen for the cis-NAT promoters (Figures B.7 and B.8) [3, 4, 25, 132]. However, histone modification levels and RNA Pol II binding are substantially more enriched around genic promoters. In addition, genic promoters show distinct enrichment patterns for H3K9Me1, H3K36Me3and H4K20Me1; these differences are likely due to the location of cis-NATs in gene bodies, which differ with respect to the distribution of these particular modifications. Overall, these results further support the functional and regulatory potential of cis-NAT promoters that are actively transcribed, albeit at lower levels than genic promoters.

It is formally possible that the cis-NAT chromatin enrichment patterns observed here can be attributed the fact that the cis-NATs were identified using CAGE, and any CAGE cluster would show such a pattern. To control for this possibility, we performed a similar analysis using CAGE clusters overlapping exons in the sense orientation, which may not be expected to show the same pattern of modification as CAGE clusters associated with genuine promoters. Indeed, sense exonic CAGE clusters have previously been suggested to represent transcriptional degradation products, as opposed to promoters, and were not found show promoter characteristic chromatin profiles [137]. Here, we performed the same set of chromatin enrichment analyses done for cis-NATs on exonic CAGE clusters. The patterns of histone modifications near exonic CAGE clusters are markedly different from those seen for cis-NAT promoters and genic promoters (Figures B.5 and B.6). These results indicate that the cis-NAT chromatin enrichment profiles observed here are not simply a generic marker for the presence of CAGE clusters.

Differential expression of cis-NAT promoters

Differential expression of cis-NATs was measured by counting the fraction of the 6 ENCODE cell types in which each cis-NAT promoter was expressed. In order to remove cis-NATs whose expression falls below the limit of CAGE detection, only those cis-NAT promoters that show activity higher than the 90th percentile in some cell type were used. On average, these cis-NAT promoters are expressed in 33% of the ENCODE cell types studied here compared to 43% seen for genic promoters (Figure 5.4a), this difference is statistically significant ($P \approx 0$, Wilcoxon rank sum) indicating that cis-NAT expression is more cell-type specific than genic expression.

Rarefaction curve analysis was used to evaluate the extent to which each individual CAGE data set uncovers novel cis-NAT promoters compared to novel genic promoters. For this analysis, the average numbers of cis-NAT or genic promoters detected across all possible CAGE data set combinations, ranging from 1-16 data sets, were calculated. Compared to genic promoters, a significantly smaller fraction of cis-NAT promoters is detected when one or only fewer than 8 CAGE data sets are considered ($P < 0.001$, Wilcoxon rank sum) (Figure 5.4b). For both genic and cis-NAT promoters, the number of new promoters detected decreases rapidly as more CAGE sets are considered, suggesting that most cis-NAT and genic promoters have been captured. The differences seen for the cis-NAT versus genic curves further underscore the extent to which cis-NATs are specifically regulated.

Association between cis-NAT and genic promoter activity

Previous studies have suggested that the presence of cis-NATs leads to the down-regulation of gene expression[44]. If cis-NATs are indeed repressive regulatory elements, then one may expect to observe a negative correlation between cis-NAT expression levels and the expression levels of the genes in which they are found. To evaluate this prediction, we regressed the activity levels of genic promoters with those of the corresponding cis-NAT promoters, however no correlation was apparent (Figures B.9 and B.10). Therefore, we used a more sensitive data mining approach to search for possible associations between genic promoter activity and cis-NAT promoter activity. To do this, genic promoters were classified as having high or low activity, and the corresponding genes were classified as having high or low cis-NAT activity in each of the 6 cell types as described above. Association mining then was used to evaluate the

levels of co-occurrence of the four possible gene and cis-NAT activity category combinations: 1) high cis-NAT & high gene, 2) high cis-NAT & low gene, 3) low cis-NAT & high gene, 4) low cis-NAT & low gene. We found that co-occurrence of high cis-NAT and high genic promoter activity occurs approximately twice as frequently as would be expected by chance (Figure 5.5 and Table B.4). Similarly, the frequency of high/low associations is much lower than would be expected and the frequency of low/low associations is higher than expected. This association remains when only those cis-NAT promoters distal (> 2.5kb downstream) to the genic promoter or proximal (< 2.5kb downstream) to the genic promoter are considered (Figures B11 and B12 and Table B.5 & B.6). These results raise the possibility that the majority of cis-NATs are activating rather than repressive regulatory elements.

Conclusions

It has been known for some time that there is active antisense transcription in the human genome, though it has only recently become appreciated how pervasive it is. However, the functional significance of human cis-NATs is a matter of debate; it is possible that many of the apparent cis-NATs actually represent transcriptional noise or degraded fragments of sequence processed from larger transcripts. Here, we have attempted to address the potential functional significance of human cis-NATs genome-wide by evaluating the chromatin environment and regulatory properties of their promoters. This approach is based on the rationale that specifically regulated promoters will have distinct chromatin profiles and protein binding properties. Accordingly, the presence and distribution of such chromatin features at the promoters of novel uncharacterized

transcripts, when considered together with their relative activity levels, can be used to provide support for their regulation and potential functional significance.

Taking advantage of methods for characterizing protein binding and histone modifications genome-wide, we demonstrate that active human cis-NAT promoters are in fact enriched for histone modifications and RNA Pol II binding. Furthermore, histone modifications and RNA Pol II binding peak at cis-NAT TSS, and the levels of histone modifications and RNA Pol II binding are correlated with the activity of the cis-NAT promoters. These data suggest that the expression of human cis-NATs is driven by RNA Pol II and at least partially regulated by the modification of histone tails. While the specific function of individual cis-NATs remains an open question, the fact that the cis-NAT promoters are bound by RNA Pol II and epigenetically modified suggests that they are specifically regulated. Indeed, the presence of both cis-NAT promoters with activating marks and cis-NAT promoters with repressive marks is consistent with the high levels of differential expression observed here for cis-NATs and tissue-specific regulation of their function. While the cis-NAT chromatin and expression features uncovered here are consistent with a functional role as regulators, they may also be taken to represent a required pre-condition of function. Definitive confirmation of the functional role for cis-NATs will await experimental validation of individual cases.

CHAPTER 6

ENDOGENOUS RETROVIRUSES AND THE EPIGENOME

Abstract

Endogenous retroviruses (ERVs) are the evolutionary remnants of retroviral germline infections, which are no longer capable of intercellular infectivity. Despite being confined within the genomes of their hosts, ERVs are able to replicate and spread via retrotransposition. This replicative process helps to ensure the elements' proliferation and long term evolutionary success, but it also imposes a substantial mutational burden on their host genomes. Accordingly, host organisms have evolved a variety of mechanisms to repress ERV transposition, including epigenetic mechanisms based on the modification of chromatin. In particular, DNA methylation and histone modifications are used to silence ERV transcription thereby mitigating their ability cause mutations via transposition. It has recently become apparent that epigenetic and chromatin based regulation of ERVs can also exert substantial regulatory effects on host genes. In this chapter, we provide a number of examples illustrating how chromatin modifications of ERV insertions relate to host gene regulation including both deleterious cases as well as exapted cases whereby epigenetically activated ERV elements provide functional utility to their host genomes via the provisioning of novel regulatory sequences. For example, we discuss ERV-derived promoter and enhancer sequences in the human genome that are epigenetically modified in a cell-type specific manner to help drive differential expression of host genes. The genomic abundance of ERVs, taken together with their proximity to host genes and their propensity to be epigenetically modified, suggest that

this kind of phenomenon may be far more common than previously imagined. Furthermore, the environmental responsiveness of epigenetic pathways suggests the possibility that ERVs, along with other classes of epigenetically modified TEs, may serve to coordinately modify host gene regulatory programs in response to environmental challenges.

Introduction

Endogenous retroviruses (ERVs) are the genomic remnants of retroviruses that integrated into a host genome and subsequently lost the ability to leave the host cell, instead replicating within the host genome [138]. Evolutionarily, ERVs are members of a broader class of mobile genetic elements known as LTR-containing retroelements; included in this broader set are the LTR retrotransposons. LTR-containing retroelements are named for the Long Terminal Repeats (LTRs) found at their 5' and 3'-ends. These LTRs are direct repeats, identical at the time of insertion, and contain regulatory sequences required for element transcription. The LTRs of ERVs and LTR retrotransposons are highly similar in structure and function [139]. The similarity between ERVs and LTR goes beyond the presence of the LTR sequences, however. In fact, LTR retrotransposons have been referred to as being 'retrovirus-like' elements due to their similarity to both ERVs and retroviruses [1]. Both ERVs and LTR retrotransposons contain coding sequences necessary for their integration into the host genome as well as a region encoding a reverse transcriptase that catalyzes the polymerization of DNA from an RNA template. Comparison of reverse transcriptase sequences from diverse retrotransposons and viruses revealed that retroviruses and ERVs are most closely grouped with LTR retrotransposons [139-141]. Phylogenetic

reconstructions based on reverse transcriptase sequence alignments indicate that retroviruses and ERVs represent a monophyletic subset of overall LTR retroelement diversity and show that the LTR retortransposons form a basal clade to this group with greater relative diversity. These data were taken to indicate that, at some time in the distant past, retroviruses emerged from within the LTR retrotransposon lineage via the acquisition of an envelope protein coding sequence that conferred intercellular infectivity, *i.e*. the ability to escape the confines of the host cell [139]. Thus, ERVs, which are a group of retrovirus-derived sequences that are no longer capable of intercellular infectivity, represent a reversion to the ancestral state of LTR retortransposons as non-infectious genomic elements.

As with other classes of retrotransposable elements, LTR-containing retroelements, including ERVs, are able to increase their copy number in the genome via retrotransposition. Through retrotransposition, LTR-containing retroelements can achieve high copy number within genomes, *e.g.* ~700,000 insertions in the human genome, comprising 8% of the total genomic sequence [1]. The retrotransposition of ERVs and other LTR retroelements can cause deleterious mutations in the host. In mouse, where ERVs are highly active, it has been estimated that 10% of *de novo* mutations result from novel ERV insertions [7, 142]. ERV insertions can cause deleterious mutations via a number of mechanisms including the induction of transcriptional aberrations in host genes. For example, integration of the Ten mouse ERV into the second intron of the Fas (tumor necrosis factor receptor superfamily, member 6) gene has been shown to lead to aberrant splicing of Fas transcripts via the donation of splice donor and acceptor sites that cause the inserted ERV to be spliced into

the nascent host gene transcript [143]. This leads to mutant mice with an autoimmune phenotype. More recently, it has been shown that insertion of a mouse ERV into to an intron of the Slc15a2 (solute carrier family 15, member 2) gene can cause pre-mature transcriptional termination at distance via a distinct mechanism that does not involve changes in the splicing of the gene [144]. This same work revealed that similar pre-maturely terminated transcripts occur in ~5% of mouse genes with intronic polymorphisms of ERVs.

In order to prevent deleterious insertions of ERVs and other LTR-containing retroelements, host genomes have evolved a variety of mechanisms to suppress element transposition [145]. Among these mechanisms, epigenetic and chromatin based silencing of insertions by the host limit the ability of the elements to produce mRNA, thereby greatly reducing the likelihood that they will be transposed [146, 147]. A number of recent studies on mammalian chromatin have demonstrated the extent to which ERV element sequences are marked with repressive histone modifications, which presumably limit their transcription. For example, using ChIP-PCR (Chromatin Immunoprecipitation followed by PCR amplification), Martens et al. demonstrated that Intracesternal A particle (IAP) insertions, a family of ERVs, are subject to the repressive H4K20Me3 (trimethylation of Histone 4 K20) histone modification, while at the same time showing very low levels of the activating mark H3K4Me3 for these same elements [148]. Similarly, using ChIP-seq (Chromatin Immune-Precipitation followed by massively parallel sequencing) [149], Mikkelsen et al. found that mouse ERVs are enriched for the epigenetically silencing histone modifications H3K9Me3 and H4K20Me3 [150]. Using

ChIP-seq data from CD4+ T-cells, Huda et al. also found that human LTR-containing retroelement insertions were enriched for silencing histone modifications [151].

While most chromatin studies of ERVs to date have focused on the epigenetic silencing of these elements for the purpose of genome defense, it has become increasingly clear that epigenetic modifications of ERVs and other LTR-containing retroelements can also have profound effects on the regulation of host genes. In other words, epigenetic modifications of ERV sequences are not only used to repress element transcription, but can also be exapted [94, 152] for the purposes of controlling host gene expression. For example, epigenetic silencing of an ERV insertion near the promoter of a host gene could possibly reduce the transcriptional activity of that gene. Alternatively, ERV or LTR-containing retroelement insertions could be actively modified and regulated in a way that benefits the host, *e.g.* as an alternative promoter for a host gene or an enhancer that regulates gene expression at distance. Such exapted insertions could help to diversify the host transcriptome as has been seen for an ERV-derived promoter driving the expression of the IL-2 receptor beta gene in human placenta [18]. In this chapter, we focus on these kinds of chromatin mediated regulatory exaptations of ERVs and other LTR-containing retroelements. We provide several examples of recent studies showing how epigenetic modifications of these kinds of elements can affect the regulation of host genes in a variety of eukaryotic species. First, we explore host gene regulatory effects exerted by the epigenetic silencing of LTR retroelements (sections 2-4), and then we focus on how activating chromatin modification of these kinds of elements can also effect the regulation of nearby host genes (sections 5-7).

Epigenetic silencing of LTR retroelement insertions in Arabidopsis thaliana

In an early study on the effect of transposable element (TE) insertions on the local chromatin environment, Lippman et al. characterized the chromatin environment of a genomic region in *Arabidopsis thaliana* which arose from an ancient segmental duplication [146]. This duplicated chromosomal region is a so-called 'knob', *i.e.* an interstitial heterochromatic region, which was found to contain many LTR retrotransposon and other TE insertions that are not present in its duplicated counterpart. These TE insertions are evolutionarily young indicating that they were inserted into the knob region after the ancient duplication by which it was generated (Figure 6.1). The coincidence of heterochromatin and novel TE insertions in the knob region was taken to suggest that these insertions led to the formation of interstitial heterochromatin after duplication, presumably as a result of host chromatin based silencing mechanisms that were targeted to these TEs. Using tiling arrays, Lippman et al. demonstrated that the TE insertions in the knob were in fact marked with DNA methylation and the repressive H3K9Me3 histone modification, with elements of the gypsy family being particularly heavily modified. Knockdown of the DNA methyltransferase *ddm1* resulted in the decrease of the levels these repressive marks in the knob region and an increase in LTR retrotransposon expression therein, mainly from the gypsy family of elements.

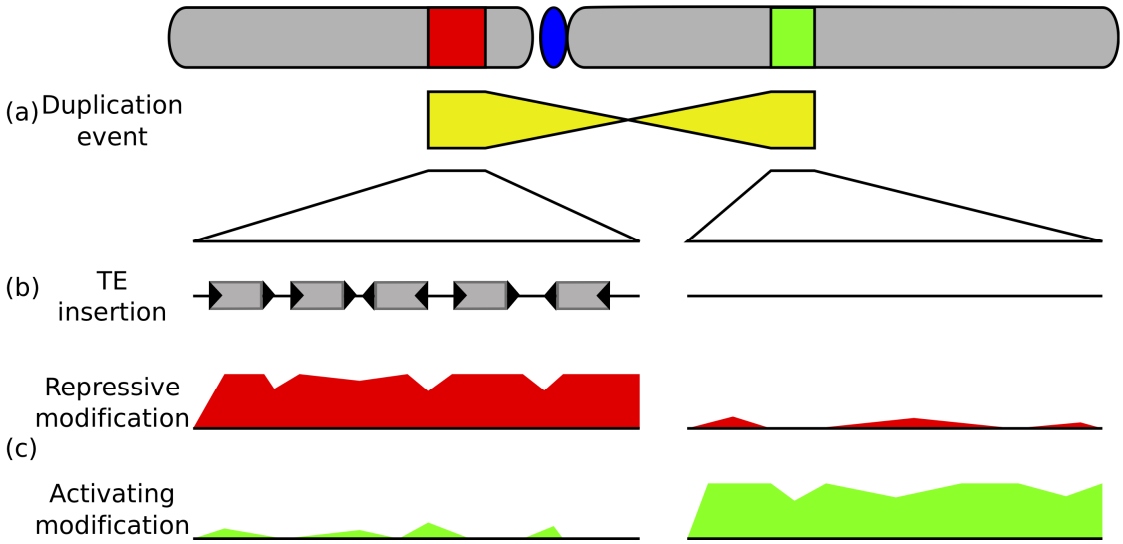


Figure 6.1. Generation of an interstitial heterochromatic region driven by transposable element (TE) insertions. (a) An ancient segmental duplication in *A. thaliana* led to two paralogous regions. (b) One of the duplicated regions is subject to multiple TE insertions (left), including numerous LTR retroelements, while the other duplicated region remains largely free of such insertions (right). (c) The region with TE insertions (left) is subject to repressive epigenetic modifications (red) and depletion of activating modifications (green), while the reverse is seen for the region without the insertions. Figure adopted from [146].

This study demonstrated that insertion of LTR-containing retroelements could lead to the in situ formation of heterochromatin in one particular region of a eukaryotic genome in response to host defense mechanisms that silence element expression. These findings suggested that the novel insertions of LTR-containing retroelements could have genome-wide effects via the generation of local heterochromatic regions that can silence nearby host genes.

Epigenetic silencing of LTR retroelement insertions and the effect on nearby genes in A. thaliana

The results from Lippman et al. demonstrated that LTR insertions generate novel heterochromatic regions in *A. thaliana*, and they also showed that genes co-located with TEs in the heterochromatic knob-region were expressed at lower levels than their paralogs located in euchromatin. Indeed, if an LTR-containing retroelement insertion near or within a transcribed locus is epigenetically silenced, then it may be possible for the element silencing to affect expression of the gene as well. Based on this line of thinking, Hollister and Gaut sought to characterize the effect of methylated TE insertions, including ERVs and other LTR-containing retroelement insertions, on the expression of nearby genes *A. thaliana* [153]. Initially, they observed a globally lower expression of genes near TE insertions; however, this did not take into account the epigenetic state of the insertion. Using genome-wide bisulfite sequencing data, they went on to demonstrate a genome-wide depletion of methylated TE insertions near genes, suggesting that such insertions are selected against, perhaps by virtue of their silencing effects on nearby gene expression. In fact, the authors demonstrated that genes proximal to such methylated insertions were expressed at lower levels, indicating that the methylation of TE insertions near genes reduces their expression. In line with the role of selection in removing methylated TEs from the proximity of genes, Hollister and Gaut demonstrated that methylated polymorphic TE insertions near genes were skewed towards rare variants. Furthermore, this effect was observed only for insertions <1.5 kb from genic loci, pointing to locally confined spreading of methylation from TE insertions into nearby or adjacent genes. Indeed, older methylated TEs were found to be farther from genes,

suggesting that selection has not acted on them as it has on younger methylated TEs near genes.

The depletion of LTR-retroelement and other TE insertions within and near genes has been observed for a number of eukaryotic species and itself strongly suggests that such insertions are selected against. The study by Hollister and Gaut provided a specific mechanistic basis for this selection, *i.e*. the fact that methylated insertions within and near genes are deleterious by virtue of their silencing effects on gene expression. Given what these authors observed, it seemed possible that the reduction of neighboring gene expression by the insertion of a TE could also occur in other species that epigenetically silence TE insertions and could perhaps be even more profound in genomes that are denser in repetitive elements.

Heterochromatin spreading from polymorphic IAP insertions in the mouse genome

The mouse IAP family of ERVs is a highly active, with ~26,000 annotated insertions [7]. While Mikkelsen et al. previously showed that IAP insertions in mouse were epigenetically silenced [150], the effect that such silencing would have on nearby genes remained largely unexplored. Recently, Rebollo et al. investigated the possibility that novel IAP insertions in mouse could lead to the formation of local heterochromatin and the spreading of heterochromatin away from the insertion into nearby sequences [154]. To do this, Rebollo et al. characterized IAP insertions which were polymorphic between two mouse cell lines, allowing them to observe the epigenetic state of the IAP insertion site with and without the insertion. It was found that the borders of IAP insertions, both those which were polymorphic between the two cell types and common IAP insertions, were enriched for the repressive H3K9Me3 histone modification. The enrichment of

H3K9Me3 was found to spread from the borders of the IAP insertion up to a maximum of 5kb. Importantly, for polymorphic IAP insertions, Rebollo et al. showed that the pre-insertion site in the cell type without the IAP insertion was not enriched for H3K9Me3, indicating that the novel IAP insertion was the source of the repressive modification.

The spreading of repressive modifications from an IAP insertion raised the question as to whether or not such spreading could lead from the insertion to a nearby promoter (Figure 6.2). Indeed, Rebollo et al. were able to find an example of a polymorphic IAP insertion proximal to a mouse gene. There is an IAP insertion upstream of the *B3galtl* promoter which is present only in the J1 cell type. In the J1 cell type, DNA methylation and the repressive histone modification H3K9Me3 extend from the IAP insertion into the promoter of the *B3galtl* gene, which is accordingly down-regulated in J1 compared to the TT2 cell line that lacks the gene proximal IAP insertion. Such a spreading of heterochromatin from LTR insertions into nearby genes, and the negative regulatory effects caused by such spreading, could explain the apparent negative selection against LTR insertions near promoters previously observed for the mouse and human genomes [68, 70].

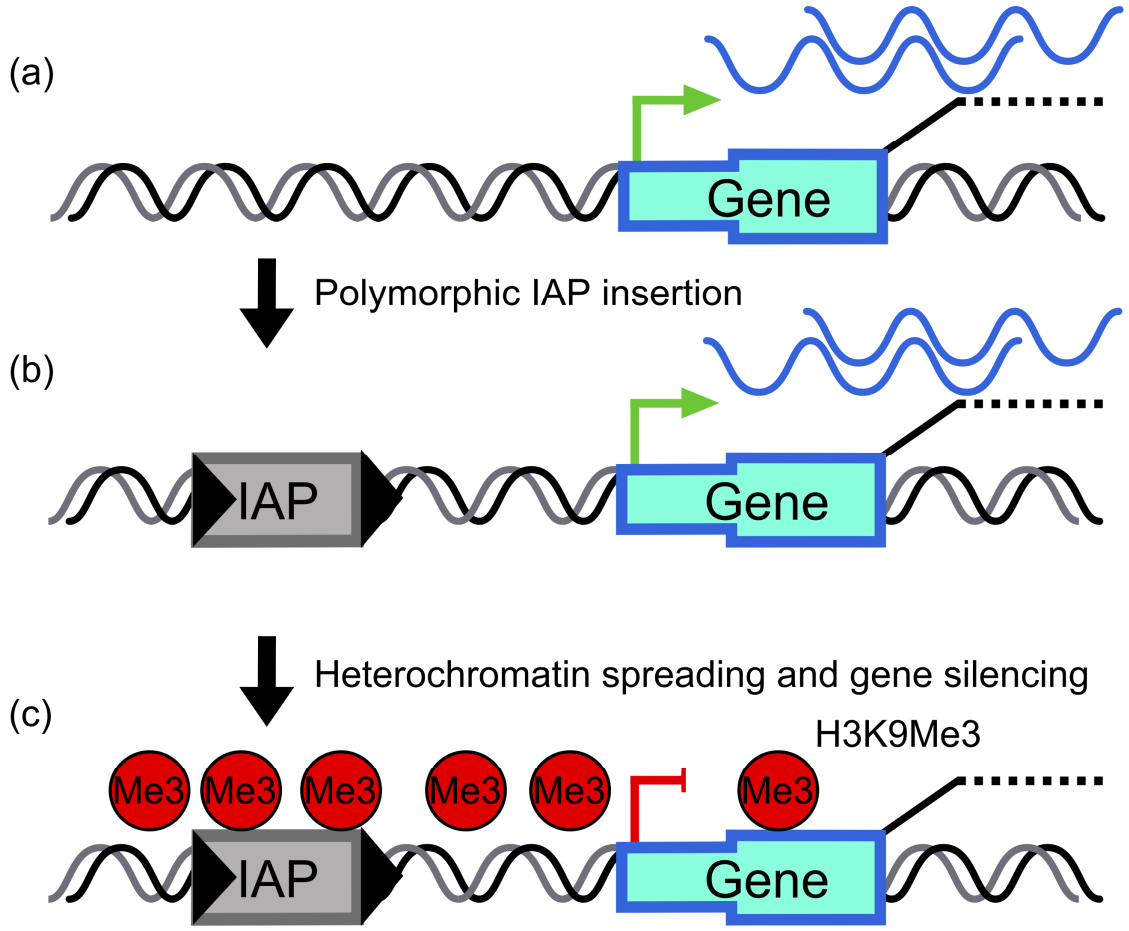


Figure 6.2. Spreading of heterochromatin from a novel IAP insertion. (a) An active mouse gene promoter region prior to an IAP insertion. (b) Cell-type specific insertion of an IAP element near the active mouse gene promoter. (c) The IAP insertion is silenced with the repressive histone modification H3K9Me3 (red circles) and this repressive mark spreads to the nearby gene promoter resulting in silencing of the gene. Figure adopted from [154].

It is worth noting that when looking for instances where the insertion of an IAP element led to heterochromatin spreading and alteration of gene expression, Rebollo et al. looked only at those IAP insertions proximal to promoters. In addition to promoters, there are many thousands of enhancers scattered within and between mammalian genes. Visel et al. characterized several thousand enhancers in mouse tissue samples, many of

which were active in only one of the cell types analyzed [155]. Similarly, Ernst et al. characterized many thousands of likely human enhances based on their profile of active histone modifications [156]. Such active histone modifications are likely important in the function of the enhancers, and it stands to reason that an IAP inserted near an enhancer could reduce its function via the spreading of repressive epigenetic histone modifications. Indeed, the insertion of an IAP element near an enhancer could conceivably affect the expression of a gene in a more specific manor than promoter proximal insertions since enhancers tend to be more cell-type specific than promoters.

Demethylation of an IAP insertion leads to ectopic expression of the agouti gene in mouse

While many ERVs are epigenetically silenced, it is likely, given the large number of insertions present in many genomes, that some will escape such silencing, or even become actively modified. Indeed, Hollister and Gaut showed that not all LTR retroelement insertions are repressed in *A. thaliana*, a large number are demethylated [153], and it would not be surprising to find that LTR retroelements in other species could also be demethylated. Given that ERVs contain their own promoters and regulatory sequences, it is conceivable that when demethylated their promoters could potentially transcribe through or away from their inserted sites into nearby genes. Given the genomic abundance of ERVs and other LTR-containing retroelements, it would seem probable that a number of demethylated insertions are likely to transcribe nearby host gene sequences. One such example of this phenomenon occurs at the *agouti* locus in mouse.

The *agouti* gene in mouse controls the pigmentation of mouse coats and hair follicle development. There exist mouse strains which show ectopic expression of the agouti gene, predisposing the mice to tumors and obesity [157]. Interestingly, the ectopic expression of the *agouti* gene is widely variable: the expression ranges from mice which express it widely, to those which show variegation in expression and those which show no ectopic expression and are otherwise phenotypically normal. It was demonstrated that the ectopic expression was not driven by the canonical promoter of the *agouti* gene, but an IAP insertion upstream of the *agouti* coding exons and that the level of expression driven from this IAP was correlated with the demethylation its LTR (Figure 6.3) [157, 158].

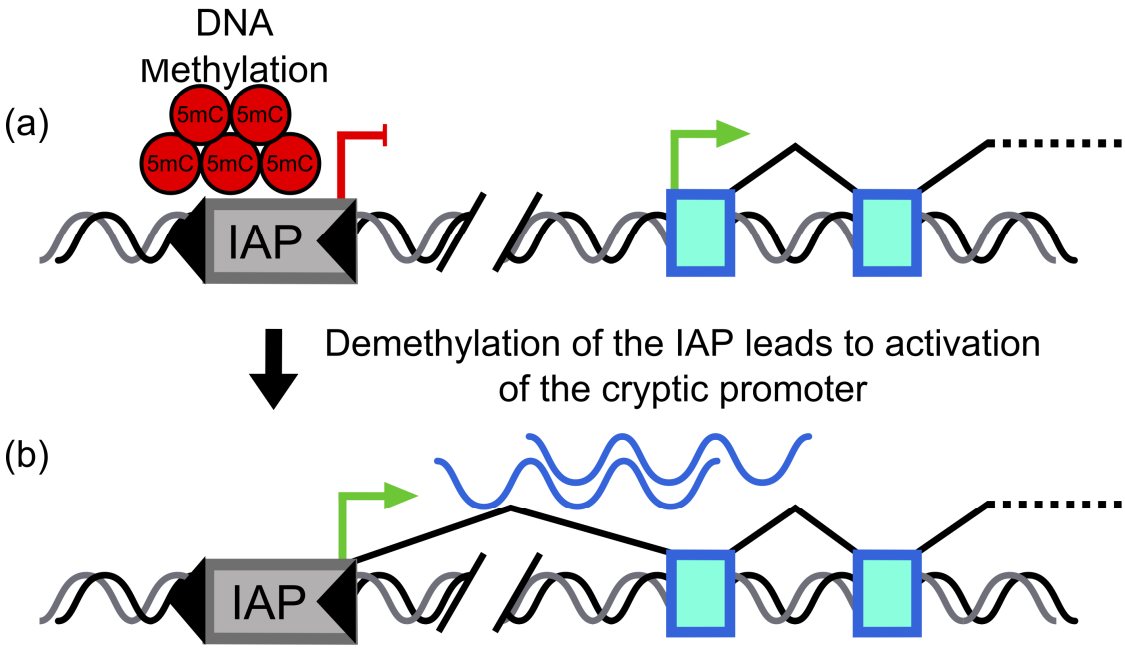


Figure 6.3. Demethylation of an IAP leads to ectopic expression of the agouti gene. (a) In phenotypically normal mice, the *agouti* proximal IAP insertion is subject to DNA methylation (5mC, red circles) and is inactive. Accordingly, agouti gene expression is driven by its canonical promoter in the appropriate tissues. (b) In mice where the IAP insertion is demethylated, it can drive ectopic expression of the nearby *agouti* gene from a cryptic promoter encoded by the IAP insertion. Figure adopted from [158].

This *agouti* locus represents a departure from the usual reasoning behind the epigenetic silencing of LTR-containing retroelements and other TE insertions: rather than preventing retrotransposition *per se*, epigenetic silencing of the IAP insertion serves to prevent deleterious transcription from the IAP insertion into the neighboring *agouti* gene. While the *agouti* case was a single example of an ERV altering genomic function when demethylated, the large number of insertions within eukaryotic genomes, ~700,000 and ~850,000 in the human and mouse genomes [1, 7], virtually guarantees that other such de-repressed LTR retroelement insertions can and do act as promoters. Further, while transcription from the IAP insertion in the *agouti* locus is deleterious, other de-repressed

insertions could prove adaptive and become exapted for function in the host genome. Indeed, several hundred promoters derived from LTR-containing retorelement insertions have been characterized in the human genome [98], the epigenetic characterization of which we discuss in the next section.

Actively modified ERVs and human gene promoters

The initial phases of the ENCODE project [25, 28] have allowed for the unprecedented characterization of the epigenetic state of the large majority of sites in the human genome, including many repetitive elements which could not previously be characterized using array based techniques. Of equal importance, the ENCODE project has allowed for the comparison of the epigenetics state between cell types. Such comparisons allow for the detection of sites with differential modification which could in turn contribute to cell-type specific patterns of gene expression. In sections 6 and 7, we review studies of host gene promoters and enhancers respectively, based on ENCODE data from human cell lines, which demonstrate activating epigenetic modifications of ERVs and other LTR-containing retroelements and show how these reactivated insertions may drive cell-type specific patterns of gene expression.

The *agouti* locus in mouse demonstrates that the insertion of an ERV insertion near a gene can lead to the use of the insertion as an alternative promoter for the gene. Indeed, ERV and other LTR-containing retroelement-derived promoters, in both mouse and human, have been characterized in several studies. A 2004 study identified 81 genes expressed in early mouse embryos for which the 5'-end, and thus the promoter, was derived from an LTR retorelement insertion [159]. A later study used Paired-End diTag (PET) data [80] to characterize 114 distinct ERV-derived promoters in the human

genome [98], and a study by Faulkner et al. analyzed a large set of CAGE (Cap Analysis of Gene Expression) [31] libraries to investigate the potential promoter activity of LTR-containing retroelement insertions in diverse human and mouse tissues [32]. While these studies characterized a breadth of LTR-containing retroelement-derived promoters, the epigenetic status and/or chromatin modifications of these insertions was not investigated.

Huda et al. investigated the epigenetic regulation of TE-derived promoters in the human genome, including those promoters derived from ERV and other LTR-containing retroelement insertions [160]. The authors identified 1,520 distinct promoters derived from TE insertions, among them over 300 promoters derived from LTR-containing retroelement insertions (Figure 6.4). Using ChIP-seq data from the GM12878 and K562 cell lines, Huda et al. characterized the epigenetic environment of the TE-derived promoters, finding an enrichment of activating modifications for active promoters along with a concomitant depletion of the sole repressive mark used, H3K27Me3. Of note, promoters derived from LTR-containing retroelements showed the greatest divergence of histone modification and activity between the GM12878 and K562 cell types. Such a divergence suggests that LTR-containing retroelement insertions have helped to diversify patterns of mammalian gene expression.

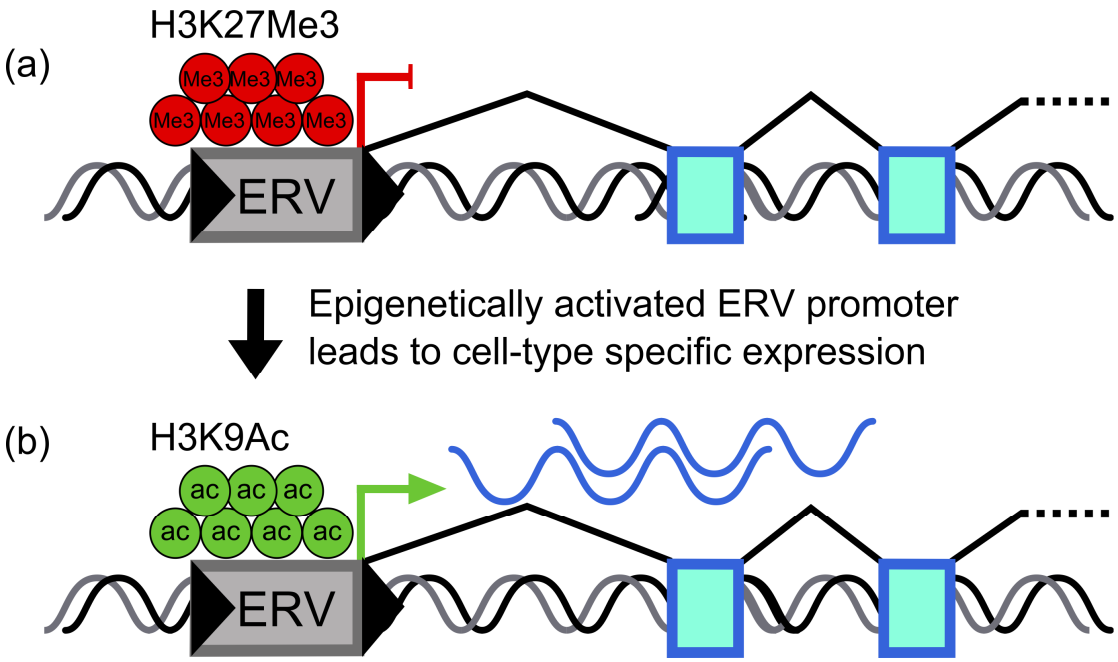


Figure 6.4. Cell-type specific epigenetic activation of human ERV-derived promoters. (a) In one cell type, a human ERV insertion is subject to repressive histone modifications and accordingly is not used as a promoter for the adjacent host gene. (b) In a different cell type, the same ERV insertion is marked with activating histone modifications, *e.g.* H3K9Ac (green circles), leading to active transcription of the adjacent host gene from the ERV promoter. Figure adopted from [160].

This study by Huda et al. demonstrated on a genome wide scale that the epigenetic activation of LTR-containing retroelement insertions can lead to the alteration of host gene expression via the use of the insertions as alternative promoters. This leads to interesting, and still largely open, questions regarding the origin and evolution of such LTR-containing retroelement-derived promoters. In the case of the *agouti* locus in mouse, ectopic transcription driven by the IAP insertion is deleterious to the mouse [157]. Given the intricate control of gene expression, one would expect that such ectopic expression would generally be deleterious. Most would therefore likely be selected against and those that can still be observed represent the few that were adaptive. Indeed,

the cell-type specific usage and epigenetic modification of the ERV and other LTR retroelement-derived promoters characterized by Huda et al. is suggestive of their adaptive nature and potential functional utility.

Actively modified ERVs and human gene enhancers

DNaseI hypersensitive sites are regions of the genome that are unusually 'open' in terms of their chromatin environment and thus susceptible to degradation by DNaseI. Such sites are often important for gene regulation, *e.g.* active promoters and enhancers. It was previously shown that a large number of DNaseI-hypersensitive sites are derived from ERVs and other LTR-containing retroelement insertions in the human genome [69], suggesting that these insertions could play roles in gene regulation apart from that of promoters, *e.g.* enhancers. Indeed, functional enhancers derived from other families of TEs are known, such as the AmnSINE1 element derived enhancers that help to drive brain specific expression [23]. Active enhancers are epigenetically modified with activating histone modifications [156, 161], and while LTR-containing retroelement insertions are typically epigenetically silenced [151], insertions acting as enhancers would be expect to show the same activating histone modifications (Figure. 6.5).

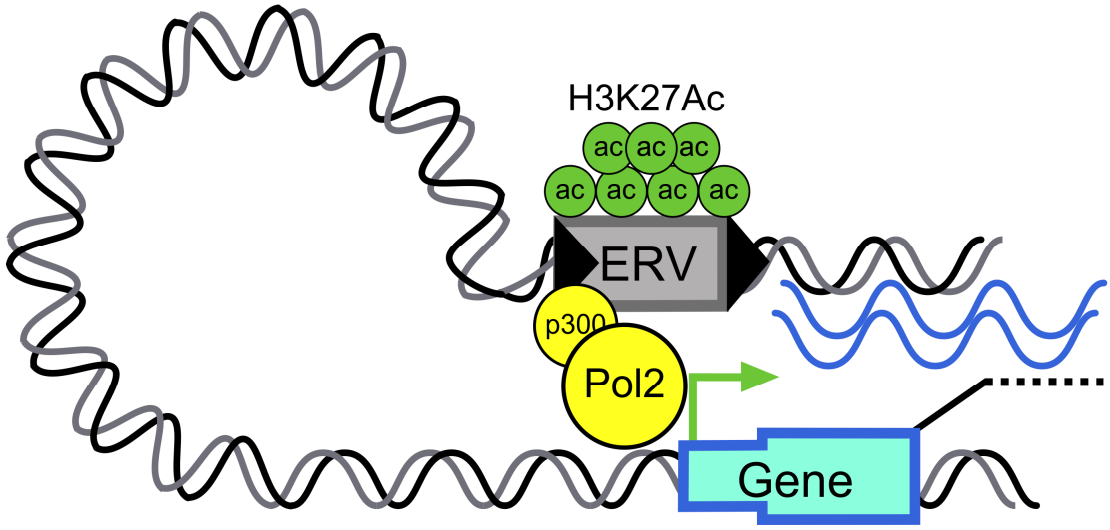


Figure 6.5. Epigenetic activation of a human ERV-derived enhancer. An ERV insertion located distal to a host gene is subject to enhancer-characteristic activating histone modifications, e.g. H3K27Ac (green circles). When activated, it acts as an enhancer for the distal gene promoter, leading to transcription from the gene promoter. Figure adopted from [162].

In a recent study, Huda et al. used the epigenetic modification patterns of enhancers to predict TE-derived enhancers on a genome-wide scale [162]. Using known p300 binding sites as a training set, the authors used ChIP-seq data from the ENCODE project in the GM12878 and K562 cell types to screen DNaseI HS sites for histone modifications similar to those of known enhancers. Nearly 20,000 such sites were identified, several thousand of which were co-located with TE insertions. Of those, over 700 sites were derived from LTR insertions. Importantly, the presence of TE enhancers correlated with the expression of nearby genes, strongly suggesting that the TE-derived enhancers characterized were active and influenced gene expression.

As in the study of TE-derived promoters by Huda et al. [160], the work on enhancers demonstrated the active epigenetic modification of human LTR-containing

retroelement insertions [162], which is in contrast with general the genome-wide enrichment of repressive modifications on such insertions [151]. Also as in the TE-promoter study, the authors used only two cell types for the analysis of TE-derived enhancers. The large majority of enhancers characterized, however, both those derived from TE insertions and other, were detected in only one of the two cell types. This is in line with what others have observed regarding the cell type specificity of enhancers. For instance, in the large scale analysis of ENCODE ChIP-seq data, Ernst et al. found that while many promoters are active across a number of cell types, the large majority of putative enhancers were active in only one of the cell types investigated [156]. This opens the possibility that there are thousands of human enhancers derived from ERVs and other LTR-containing retroelement insertions, many of which would remain unidentified in a study of only two cell-types, and underscores the potential impacting on cell-type expression of thousands of human genes that these ERV-enhancers may exert.

Conclusions and prospects

In this chapter, we reviewed some of the ways in which ERV effects on host gene regulation are mediated by epigenetic and chromatin modifications. ERVs are of course just one class of TEs, and TEs were originally discovered by Barbara McClintock by virtue of the regulatory effects they exert on maize host genes [163]. In light of these effects, McClintock referred to TEs as controlling elements, and she ultimately came to believe that TEs could actually re-organize genomes in response to environmental challenges [164]. For McClintock, this genome reorganize process was related to the genome dynamics of TEs per se, *i.e.* their ability to transpose and cause genomic rearrangements. Here, we would like to pose the idea that the TE-mediated

environmental responsiveness of eukaryotic genomes may also be attributed the epigenetic and chromatin based regulatory effects that they exert on host genes. This notion is based in part on observations that epigenetic changes can in fact occur in response to environmental stimuli [165]. In the case of ERVs, environmentally programmed ERV-mediated chromatin based regulatory changes have been observed for the agouti locus where environmental exposure to methyl donors leads to increased repression of the upstream IAP thereby mitigating the mutation ectopic expression phenotype [166]. Given the abundance of ERVs, their widespread genomic distribution and proximity to genes along with their propensity to be epigenetically modified, these elements may provide a means for host genomes to mount dynamic epigenetically programmed responses to environmental challenges.

CHAPTER 7

CELL TYPE-SPECIFIC TRANSCRIPTION TERMINATION BY TRANSPOSABLE ELEMENT SEQUENCES

Abstract

Background

Transposable elements (TEs) encode sequences necessary for their own transposition, including signals required for the termination of transcription. TE sequences within the introns of human genes show an antisense orientation bias, which has been proposed to reflect selection against TE sequences in the sense orientation owing to their ability to terminate the transcription of host gene transcripts. While there is evidence in support of this model for some elements, the extent to which TE sequences actually terminate transcription of human gene across the genome remains an open question.

Results

Using high-through sequencing data, we have characterized over 9,000 distinct TE-derived sequences that provide transcription termination sites for 5,747 human genes across eight different cell types. Rarefaction curve analysis suggests that there may be twice as many TE-derived termination sites (TE-TTS) genome-wide among all human cell types. The local chromatin environment for these TE-TTS is similar to that seen for 3' UTR annotated (canonical) TTS and distinct from the chromatin environment of other intragenic TE sequences. However, those TE-TTS located within the introns of human genes were found to be far more cell type-specific than the canonical TTS. TE-TTS were much more likely to be found in the sense orientation than other intragenic TE sequences

of the same TE family, and TE-TTS in the sense orientation terminate transcription more efficiently than those found in the antisense orientation. Alu sequences were found to provide a large number of relatively weak TTS, whereas LTR elements provided a smaller number of much stronger TTS.

Conclusions

TE sequences provide numerous termination sites to human genes, and TE-derived TTS are particularly cell type-specific. Thus, TE sequences provide a powerful mechanism for the diversification of transcriptional profiles between cell types and among evolutionary lineages, since most TE-TTS are evolutionarily young. The extent of transcription termination by TEs seen here, along with the preference for sense oriented TE insertions to provide TTS, provide an explanation for the observed antisense orientation bias of human TEs.

Background

Different kinds of somatic cells in within an individual human contain the same genome, but are obviously functionally distinct. Thus, the cell type-specific regulation of the genome, rather than the genome itself, defines the characteristics of a cell type. The importance of cell type-specific activity of promoters in the function of different cell types has long been appreciated; however, the role of cell type-specific termination of transcription has not been as well studied. Nevertheless, cell type-specific variation in transcription termination, primarily via 3'UTR shortening, has been shown to be important in cancer and in other proliferating cells [39, 41, 167]. For this study, we explored the idea that transposable element (TE) sequences, many of which can terminate

transcription, may play important roles in the cell type-specific termination of transcription.

There are numerous transposable element (TE) derived sequences in the human genome, comprising more than two-thirds of the total sequence [1] and many of these TEs are located within the introns of human genes. TEs contain their own regulatory sequences, including specific signals which lead to the termination of transcripts initiated from element promoters; human endogenous retroviral elements (HERVs), for example, have polyadenylation signals in their long terminal repeat (LTR) regions that terminate transcription [168]. Thus, numerous TE sequences located within, or nearby, human gene sequences may contribute substantially to the termination of gene transcription via the provisioning of termination signals.

There are several known examples whereby TE sequences located within, or nearby, human genes have been shown to terminate transcription of genic mRNAs. An early study of HERVs provided the first direct evidence that TE-derived sequences can terminate the transcription of non-TE human mRNAs and further suggested that different subfamilies of these elements may serve to terminate transcription in a cell type-specific manner [20]. Later, the same family of ERVs was demonstrated to terminate transcription of a novel alternatively spliced version of the human NAAA gene [169]. There is also experimental evidence showing that L1 (LINE) retrotransposon sequences can terminate the transcription of human genes, and in this same study the intronic content of L1 sequences in human genes was found to be negatively correlated with their expression levels [14]. A later study showed a similar trend whereby the presence of

polymorphic L1 insertions in human genes was correlated with a decrease in their expression in a tissue-specific manner [170].

Despite the evidence cited above indicating that TE sequences can terminate transcription of human genes in a cell type-specific manner in some cases, the extent of this phenomenon and its overall effect on the human genome sequence and cell type-specific transcriptomes have not been fully explored. Interestingly, there is a strong antisense orientation bias for TE sequences within human genes [167, 171], and this observation has been attributed to the propensity of TEs to terminate transcription when inserted in the same (sense) orientation as gene transcription [70]. Presumably, many such sense-oriented insertions would be selected against, owing to their termination of human gene transcripts, leaving a relative bias of antisense TE insertions. Consistent with this implied genome-wide effect of TE sequences on the termination of human gene transcripts, a pair of recent genome-scale surveys of transcription termination by TEs revealed ~3,000 cases of human transcripts that terminate with TEs [172, 173]. These studies, while intriguing, relied on relatively low throughput transcriptomic technologies and were not able to address cell type-specificity of TE transcription termination. Thus, the full extent of TE transcription termination within the human genome, and equally as important the cell type-specificity of this phenomenon, remains unknown.

Here, we deeply interrogated the contribution of TE sequences to human gene transcription termination via the integrated analysis of high-throughput transcriptomic data and TE-gene annotations. Since TE sequences have been shown to contribute disproportionately to cell type-specific regulation [160], we also evaluated the extent to which that transcription termination of human genes by TEs is cell type-specific. To do

this, we have characterized the space of transcription termination sites (TTS) derived from TE insertions in eight different ENCODE cell types. For these TE-TTS, we characterized the contributions from different TE families, as well as their relative insertion orientations. We found 9,287 TE-derived sequences that terminate the transcription of 5,747 human genes. Our results also show that TEs terminate transcript much more efficiently when inserted in the sense orientation relative to gene transcription and thus lend credence to the previously articulated notion that TE orientation biases result from selection against TE termination of gene transcription. We also show that TE termination of gene transcription is highly cell-type specific and thus may contribute to the specialization of cellular function through differential gene regulation.

Methods

Characterization of transcription termination sites (TTS)

Mappings of ENCODE PET data from the GM12878, H1HESC , HepG2, HeLaS3, HUVEC, K562, NHEK and prostate cell types were downloaded from the ENCODE repository on the UCSC genome browser for the hg18 version of the human genome [28, 80]. PET data from nucleus (GM12878, HepG2, HeLaS3, HUVEC, K562 and NHEK) or whole-cell (H1HESC and prostate) were used to characterize TTS. PET 3'-ends from the same data set that overlapped or were separated by 20 or fewer bases were taken as putative TTS clusters. Only those TTS clusters that had a normalized PET tag count of at least 20 per ten million tags mapped in at least one cell type were considered for further analysis. For these clusters, the specific locations of the TTS for each cluster were taken to be the base with the highest density of mapped PET 3'-ends.

TTS clusters across different cell types that overlapped by at least 80% were taken to be the same TTS.

UCSC gene model annotations [174] were used to associate TTS defined in this way with known human genes. A TTS was considered to be associated with a gene if the linked 5' ends of the PET tags were mapped to the annotated promoter of the gene and the linked 3' end TTS cluster was found within the annotated transcriptional united or up to 5kb downstream of the canonical annotated TTS. Human gene TTS characterized in this way were then co-located with TE sequences using the RepeatMasker annotations [60]. As it has been previously shown that transcription termination occurs within 50bp of the polyadenylation signal [175], TE-TTS were defined as those TTS clusters for which the peak base was at least 50bp downstream from the start of a TE insertion and less than 15bp downstream of the end of the insertion.

Histone modification enrichment analysis

The chromatin environment of PET-characterized TTS was characterized using ENCODE ChIP-seq data [156]. Where available for the same cell types as the PET data, ChIP-seq reads for the H3K9Ac, H3K27Me3 and H3K36Me3 modifications were downloaded from the ENCODE repository on the UCSC genome browser [28, 29] were mapped to the human genome reference sequence (UCSC hg18; NCBI build 36.1) using the Bowtie short read alignment utility [110]. Tags which mapped to multiple locations were resolved using the GibbsAM utility [134]. The average numbers of ChIP-seq tags were found in 5 base-pair windows +/- 5kb of (1) TE-derived TTS (2) intragenic TE insertions that do not provide a TTS and (3) non-TE derived TTS.

Utilization of PET-characterized TTS

TTS for which the region including the TTS had a normalized PET tag count of at least 20 were designated transcribed. From each cell type, those regions which were in the top 75% most transcribed, as calculated using PET tag counts, in that cell type were designated as actively transcribed. For both TE-TTS and non TE-TTS, the utilization of actively transcribed TTS in a cell type was determined by first determining the number of PET tags which begin upstream of the TTS, and which terminate in the TTS or downstream of the TTS. The utilization was then calculated using the following formula:

$$utilization = \frac{reads\ terminated}{reads\ terminated + reads\ passing}$$

Cell type-specificity of TE-derived TTS

A TTS was considered for differential utilization if the TTS (1) had a strength of utilization of at least 20% in at least one cell type and the region was (2) actively transcribed (as described above) in at least 3 cell types. The cell type specificity of a given TE-TTS was calculated using the following formula:

$$cell\ type - specificity = \frac{\sum_{i=1}^{cell\ types-1} MAX(utilization) - utilization_i}{cell\ types - 1} \Big/ MAX(utilization)$$

Estimation of the total number of TE-TTS and genes with TE-TTS

To estimate the upper bound for the number of TE-derived TTS in the human genome, we found, for all possible combinations of the eight cell types used here, the number of TE-derived TTS found with each combination. A logarithmic trend line was used to estimate the number of TE-derived TTS that would be found with increasing

numbers of cell types. The same analysis was applied for the total number of human genes that bear at least one TE-TTS.

Results and Discussion

Characterization of transposable element-derived termination sites

We characterized TE sequences that provide transcription termination sites (TE-TTS) to human genes using Paired-end diTag (PET) data. PET is a technique for the high-throughput characterization of the 5' and 3' ends of mature full-length mRNAs [80], which allows for deep annotation of paired transcription start (TSS) and termination sites (TTS) including the discovery of many novel alternative sites. TE-TTS were characterized by co-locating TE sequences with 3' PET tag clusters that are paired with 5' PET tag clusters mapped to known human gene promoters (see Methods, Table C.1). Using PET data from eight different ENCODE cell types (GM12878, H1HESC, HeLaS3, HepG2, HUVEC, K562, NHEK and Prostate) [28, 29], we discovered 9,287 distinct TE-TTS that terminate the transcription of human genes. PET data from these cells revealed a total of 89,345 non TE-derived TTS, including canonical previously annotated TTS. Overall, 9.4% of human gene TTS are provided by TE-derived sequences, and 28% of human gene loci have at least one TE-TTS.

The breakdown of TSS contributed by different TE families and the locations of these TE-TTS within human gene loci are shown in Table 7.1. While many TE-TTS correspond to the 3'UTRs and canonical TTS of human genes (21%), the majority of TE-TTS represents alternative TTS found within gene boundaries (70%) and yield creating truncated transcripts. A small minority of these alternate TE-TTS (8%) is found within upstream of coding sequences, representing messages that are severely truncated or aborted albeit in a site-specific and reproducible manner. TE-sequences also provide TTS downstream of the canonical TTS of human genes (8%) providing longer alternative transcripts. Overall, 87% of the total TE-TTS locations correspond to alternative TTS

compared to 81% for non TE-TTS, indicating that TE sequences are utilized as alternative terminators at a slightly higher frequency than non TE-sequences.

Table 7.1. Locations of human gene transcription termination sites (TTS) characterized using PET data.

TTS Location[a]	TE Family[b]									
	Non TE	**All TE**	**Alu**	**ERV**	**hAT**	**L1**	**L2**	**MaLR**	**MIR**	**TcMar**
5'UTR	3,677	704	351	52	34	114	57	17	55	22
Internal	46,716	5,404	2,975	162	334	844	377	110	381	159
3'UTR	15,491	733	267	25	69	120	62	29	102	44
Canonical[c]	16,031	1,152	293	109	102	231	124	67	151	59
Downstream[d]	2,806	673	224	70	51	142	49	45	54	33
Sum	84,721	8,666	4,110	418	590	1,451	669	268	743	317

[a] The locations of TTS characterized using PET data from eight ENCODE cell types were characterized relative to known human gene models
[b] TTS genic locations are shown for non TE-TTS and for the top eight TTS contributing TE families.
[c] TTS located within 250bp of canonical TTS
[d] TTS located up to 5kb downstream of previously canonical TTS

Several examples of human genes with TE-TTS are show in Figure 7.1. Transcription initiated from the GALNT2 promoter can terminate within an ERVL insertion in the first intron of the locus or in two canonical TTS in the 3' UTR (Figure 7.1a). TE-derived termination of GALNT2 occurs in a cell type-specific manor; most GALNT2 transcripts (78%) utilize the ERVL-derived TTS in the GM12878 cell type, whereas virtually all GALNT2 transcripts read through the ERVL insertion in the NHEK cell type and instead utilize the two canonical TTS. The transcript resulting from

utilization of the ERVL-derived TTS is severely truncated and therefore highly unlikely to produce a functional protein. Thus, while this gene is transcribed at high levels in both cell types, the ERVL-derived terminator serves to effectively reduce GALNT2 expression in GM12878 compared to NHEK. Similarly, EPHX2 transcription can terminate within an AluJb insertion in the sixth intron, resulting in a truncated transcript (Figure 7.1b). This termination is also cell type-specific, with the majority of transcripts (66%) utilizing the AluJb-derived TTS in the K562 cell type and a minority (24%) GM12878.

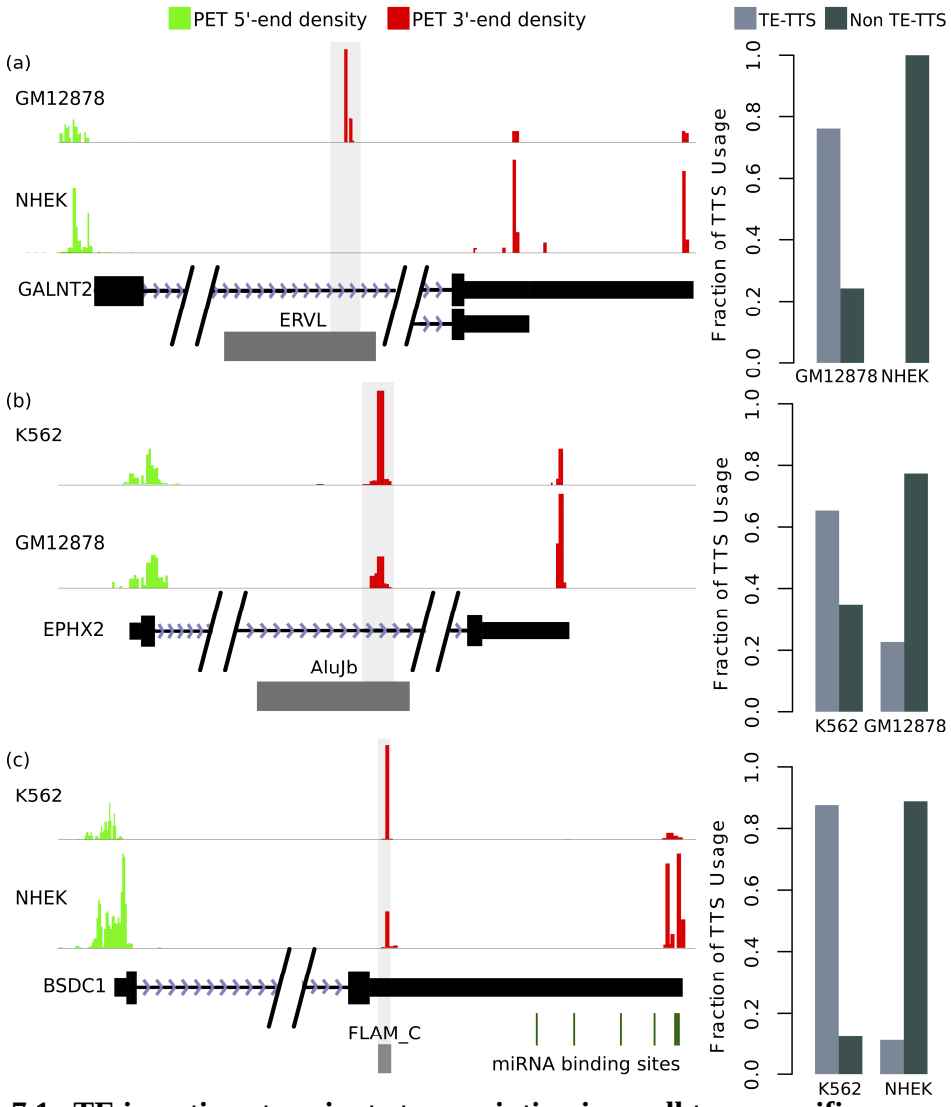


Figure 7.1. TE insertions terminate transcription in a cell type-specific manor. Clusters of linked paired-end ditag (PET) sequences that indicate the locations of the 5' (green) and 3' (red) ends of full-length transcripts expressed in different cell-types are shown above gene models indicating the locations of exons, introns and TEs that terminate transcription. For each example, the cell-type specific fractions of TTS usage for TE-TTS and non TE-TTS are shown. (a) An ERVL insertion within the first intron of the GALNT2 gene terminates the majority of transcripts in the GM12878 cell type, with a small number terminating in the two canonical TTS. No transcripts terminate within the ERVL in the NHEK cell type. (b) An AluJb insertion within the seventh intron of the EPHX2 gene terminates the majority of transcripts in the K562 cell type, while the majority of transcripts read through this sequence in the GM12878 cell type. (c) Termination of transcription within a FLAM_C insertion in the 3'UTR of the BSDC1 gene results in a shortened 3'UTR. The FLAM_C-derived TTS is utilized extensively in the K562 cell type while the majority of transcripts read through this sequence in the NHEK cell type.

Though many alternative TE-derived TTS occur within an intron of a coding locus as seen for GALNT2 and EPHX2, a substantial fraction (8.5%) occurs within a 3'UTR. For example, a TTS derived from a FLAM_C TE sequence in the BSDC1 gene is found at an alternative upstream position in the 3' UTR (Figure 7.1c). The canonical BSDC1 TTS is found several kb downstream of the TE-TTS and results in an unusually long (3.2 kb) 3'UTR. Transcripts with these kinds of long 3'UTRs are more likely to be degraded via nonsense-mediated decay [176, 177], and the annotated 3'UTR of BSDC1 also contains 10 miRNA binding sites which could be used to degrade the mRNA or reduce its translation. Thus, utilization of the FLAM_C-derived TTS, which would generate a transcript with a full-length ORF but a drastically shortened 3'UTR (~300 bp) lacking miRNA binding sites, could effectively increase expression of BSDC1 by evading post-transcriptional degradation via nonsense-mediated decay and/or miRNA binding. As is the case for the GALNT2 and EPHX2 genes, the utilization of this TE-TTS is cell type-specific, with the majority of transcripts in K562 utilizing the FLAM_C derived TTS and the majority reading through the TE-TTS in NHEK cells (Figure 7.1c). The contribution of TE sequences to alternative transcription termination is further explored later in the manuscript.

In an effort to further characterize the TE-TTS discovered here, we used ENCODE ChIP-seq data for the locations of histone modifications [28, 29, 150] to evaluate their local chromatin environment. We found that the histone modification signatures of TE-TTS are generally similar to those of non TE-TTS and distinct from intragenic TE insertions that do not provide a TTS. Different histone modifications showed distinct patterns of enrichment near TTS, and we show representative examples

of TTS histone modification signatures for an active transcriptional mark (H3K9Ac), a mark of transcriptional elongation and gene boundaries (H3K36Me) and a repressive mark (H3K27Me3) in the K562 cell type. H3K9Ac shows a marked peak of enrichment upstream of both TE-TTS and non TE-TTS and then the levels fall off precipitously after the TSS (Figure 7.2a-c). H3K27Me3 shows a slight increase downstream of the TTS for non TE-TTS, however the enrichment level was generally very low (~.1 tags per million mapped). This downstream increase in H3K27Me3 was not seen for the TE-TTS (Figure 7.2d-f), though this could be due to the comparatively low number of TE-TTS compared to non TE-TTS together with the relatively low number of H3K27Me3 marks seen within actively transcribed genes. The H3K36Me3 modification shows a more symmetrical distribution around TTS with peaks for both TE-TTS and non TE-TTS compared to intragenic TEs that do not show TSS related peaks (Figure 7.2g-i). Qualitatively similar results were seen in the GM12878 and NHEK cell types (Figures C.1-C.2). Overall, the similar local chromatin environments seen for TE-TTS and non TE-TTS suggest that the TE-TTS characterized represent *bona fide* terminators as opposed to transcriptional noise.

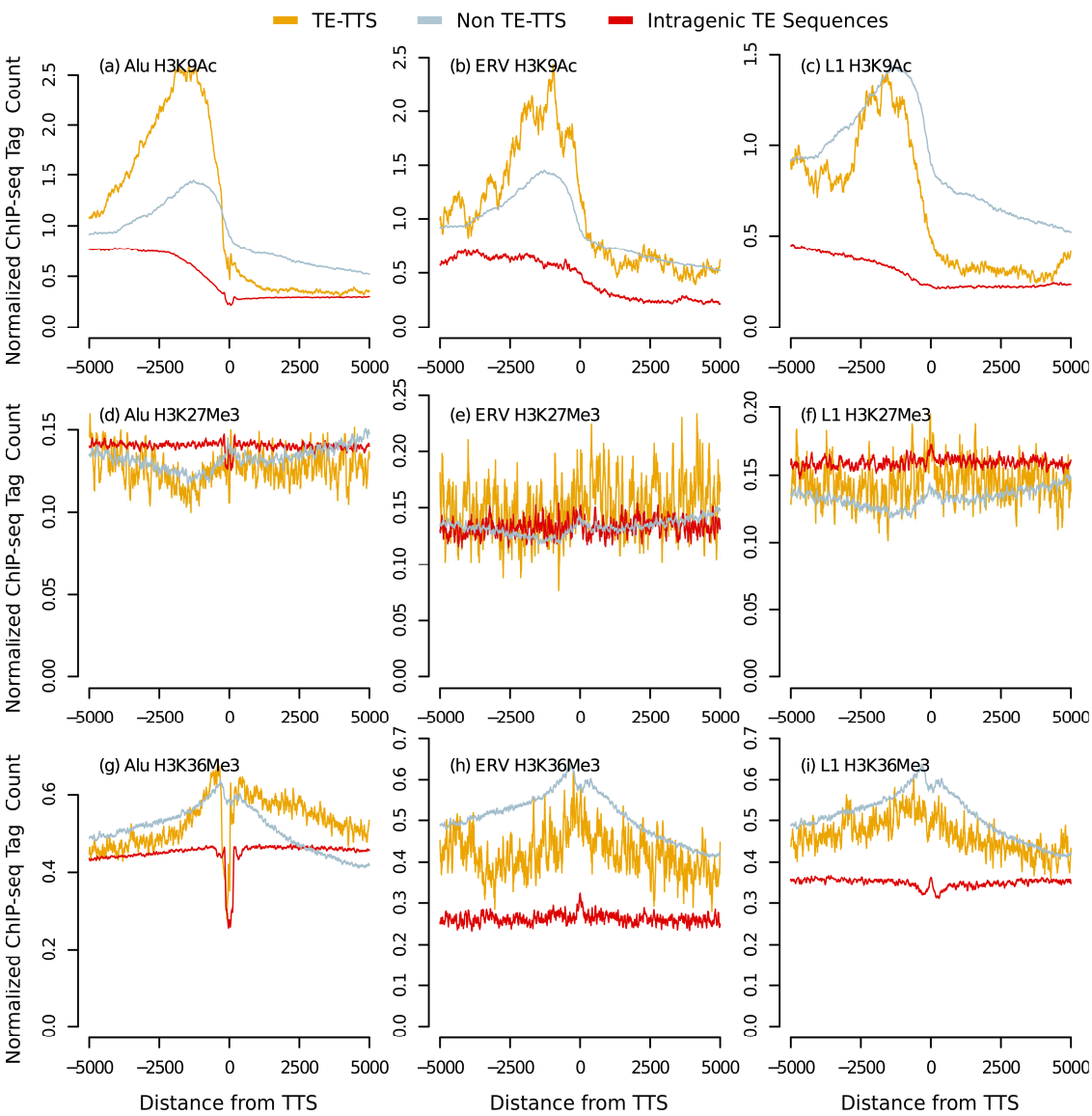


Figure 7.2. The chromatin environment of TE-TTS is similar to that of non-TE-TTS and distinct from intragenic TE sequences that do not terminate transcription. The locations of TTS and the ChIP-seq tag counts corresponding to H3K9Ac (a-c), H3K27Me3 (d-f) and H3K36Me3 (g-i) are shown for the K562 cell type. Enrichment curves, showing the average normalized numbers of ChIP-seq tags in 10 base-pair windows +/-5kb of TE-TTS (orange), non TE-TTS (gray) and intragenic TE sequences that do not show a TTS (red), are shown for three TE families, Alu (a,d,g), ERV (b,e,h) and L1 (,c,f,i).

TE transcriptional termination and insertion orientation bias

The vast majority of TE sequences within human genes are found in the antisense orientation relative to the direction of transcription of the gene [171]. The genic orientation bias of human TEs is thought to reflect differential selective elimination of sense TE insertions over time rather than a preference in the introduction of antisense insertions at the moment of transposition. The ability of TEs to cause premature termination of gene transcripts, thereby reducing levels of transcription, has been proposed as a mechanism to explain the selective elimination of sense oriented L1 sequences from human gene loci [14]. In order to investigate the role of TE-TTS in the selection against sense oriented TE insertions genome-wide, we compared the insertion orientations of intragenic TEs that do not provide TTS versus the orientations of TE-TTS for the eight largest families of human TEs (Alu, ERV, hAT, L1, L2, MaLR, MIR, and TcMar).

Seven out of eight TE families show the expected antisense orientation bias for intragenic TE insertions for which there is no evidence of TTS activity (Figure 7.3). In other words, since these antisense TE insertions do not terminate transcription, their presence within human genes is tolerated by selection. The LTR element families, the ERVs and MaLRs, show the strongest antisense orientation bias with intragenic insertions being found in the antisense orientation twice as often as the sense orientation. Conversely, Alu insertions show a much weaker antisense orientation bias. The relatively stronger bias seen for LTR element insertions suggests the possibility that there is stronger selection against sense LTR insertions and that such sense LTR element

insertions may be more deleterious. This point is explored in more detail later in the manuscript.

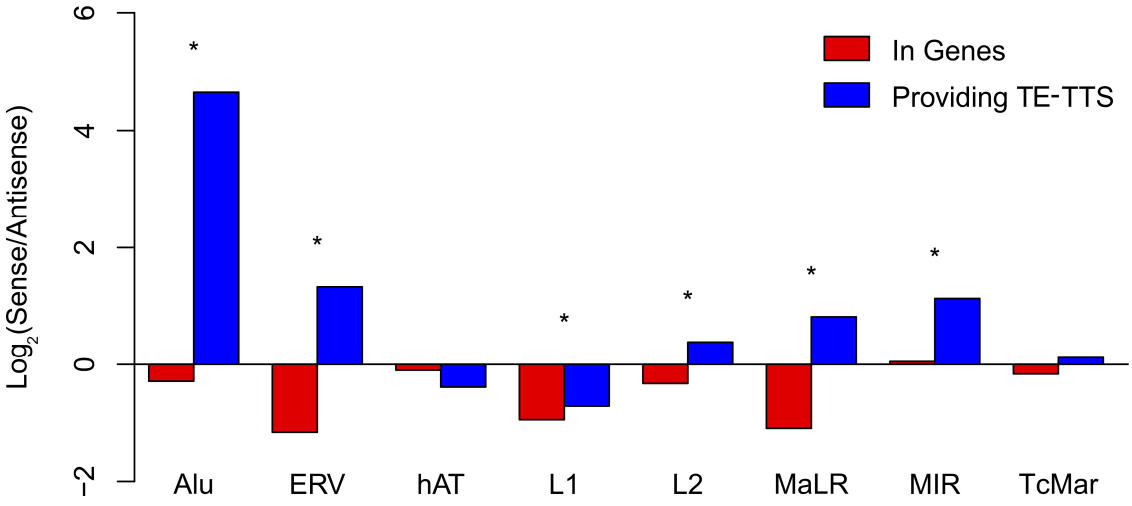


Figure 7.3. TE sequences providing transcription termination sites show a strong sense bias. For each TE family, the sense/anti-sense ratio was determined for all intragenic insertion (red) and only for those TEs that provide a TE-TTS (blue). For each TE family, statistical significance levels for the differences in the sense/anti-sense rations (* indicates P<.005) was determined using a Chi-squared distribution with *df*=1.

For those genic TE sequences that provide a TTS, the majority of TE families show a significant enrichment of insertions in the sense orientation versus the other insertions. Alus have one of the weaker antisense orientation biases for genic elements, but Alu-derived TTS show far and away the strongest sense bias; an Alu insertion providing a TTS is approximately 20x more likely to be in the sense orientation than the antisense orientation. While LTR element genic insertions show the strongest overall antisense bias, insertions providing a TTS are also much more likely to be in the sense orientation; an LTR element providing a TTS is 4x more likely to be found in the sense

orientation than the average genic LTR element insertion. The strong sense orientation enrichment seen for TE-TTS indicates that genic TEs oriented in the same direction as transcription are much more effective transcription terminators, consistent with the notion that sense oriented TE insertions are selected against owing to their disruptive effects on gene expression.

The only exception to this pattern is seen for the relatively ancient family of MIR TEs. MIRs have previously been implicated as providing gene regulatory sequences in a number of studies, and the MIR sequences that remain intact and recognizable in the human genome are likely to have been conserved by purifying selection [162]. Thus, the lack of orientation bias for MIRs, irrespective of their status as TTS, may reflect their general utility as gene regulators, rather than an ephemeral presence as neutral sequences that will be eventually lost by mutational decay.

Contributions of Alus to transcriptional termination

Given the diversity of TE insertions found in-and-around human genes, we sought to characterize the relative TE-TTS contributions of the eight largest families of human TEs (Alu, ERV, hAT, L1, L2, MaLR, MIR, and TcMar). To do this, we compared the observed numbers of TE-TTS for the different families to the expected numbers based on their genic frequencies (Figure 7.4). While L1s contribute the most TE sequence genome-wide, Alus are the most abundant genic TE family (31% of all genic TE insertions) (RepeatMasker). Thus, Alu insertions would be expected to provide a large number of TE-derived TTS. However, previous studies have characterized ~400 Alu insertions providing TTS, a substantially smaller than expected fraction [172, 173]. In contrast to these findings, we found that Alu-TTS were more abundant than TTS derived

from other TE families, providing 43% percent (4,551) of all TE-TTS, far more than would be expected based on the frequency of Alu genic insertions. Other TE families generally contributed fewer TTS than expected based on their genic frequencies, with MIR-derived TTS being far less common than expected; ERV was the only other TE family to provide significantly more TTS than expected.

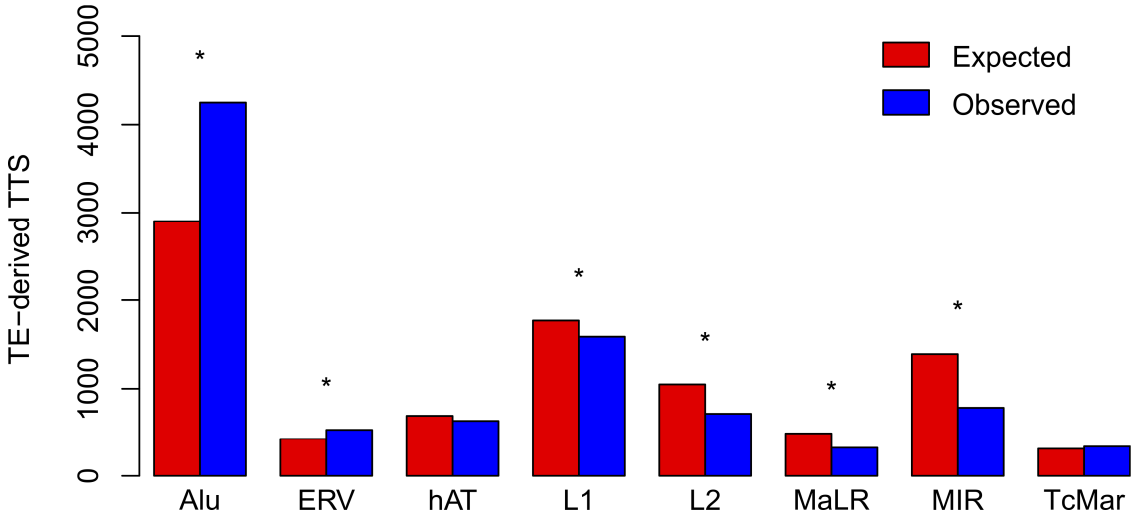


Figure 7.4. Alu family sequences provide a greater than expected number of TTS. Expected (red) versus observed (blue) counts of TTS derived from different TE families. Expected counts of TTS derived from each TE family were calculated based on the fraction of intragenic sequences. For each TE family, statistical significance levels for the differences between the expected versus observed counts (* indicates $P<10^{-5}$) were determined using a Chi-squared distribution with *df*=1.

The over-abundance of Alu-TTS could be attributed to their functional utility as expression regulators, or it could simply reflect the fact that Alu-TTS are not as disruptive and therefore more tolerated by selection. Consistent with the latter neutral scenario, the over-abundance of Alu-TTS may reflect their relatively young age, suggesting that there has not been adequate time for their removal from the genome. To

evaluate these possibilities, we evaluated the TTS contributions of Alu subfamilies of different ages (FLAM, AluJ, AluS and AluY). Relatively older Alu subfamilies (FLAM & AluJ) contribute more TTS than expected, whereas the younger subfamilies (AluS & AluY) contribute fewer than expected (Figure 7.5a). For instance, even though FLAM elements are found in less than half the genic frequency of AluY insertions, they contribute more TTS to human genes. These observations argue against the neutral explanation for the abundance of Alu-TTS. To explore this further, we evaluated the strength of utilization for TTS derived from the different Alu subfamilies. The strength of utilization for any TTS is measured as the relative frequency with which it terminates transcription versus the frequency that it is read through (see Methods). Consistent with what is seen for the relative levels of TTS donation by the different Alu subfamilies, older families show higher levels of TTS utilization than do younger families (Figure 7.5b), suggesting the possibility that many of these Alu-TTS are preserved via selection by virtue of their functional utility for the host gene.

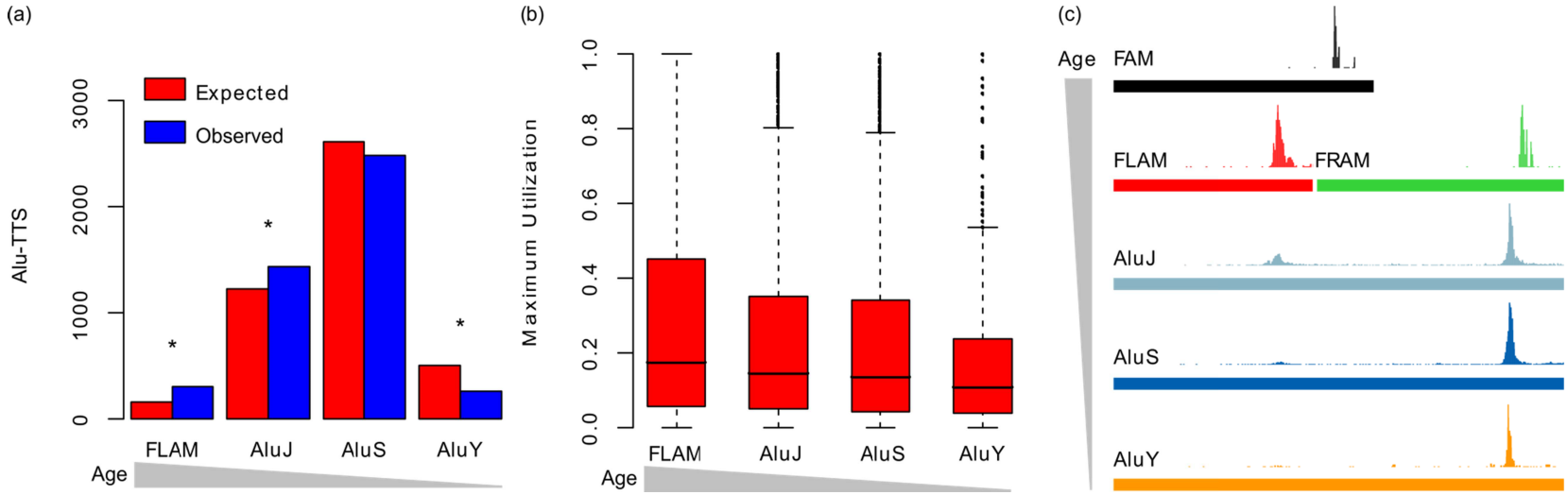


Figure 7.5. Alu-TTS are not randomly distributed in Alu insertions and older Alu families are over-represented. (a) Expected (red) versus observed (blue) counts of Alu-TTS are shown for individual subfamilies of different ages (older-left to younger-right). Expected counts of TTS derived from each subfamily were calculated based on the fraction of intragenic sequences. For each Alu subfamily, statistical significance levels for the differences between the expected versus observed counts (* indicates $P<10^{-4}$) were determined using a Chi-squared distribution with *df*=1. (b) Distributions of maximum utilization values (see Methods) for Alu-TTS are shown for individual subfamilies of different ages (older-left to younger-right). (c) For Alu-TTS provided by elements of different subfamilies, the position of each TTS within the subfamily consensus sequence was determined and the density of all TTS along the length of each subfamily consensus sequence is indicated by the height of the peaks.

In light of the exceptional ability of Alus to provide TTS to human genes, we explored the specific sequence context by which these elements terminate transcription. To do this, we mapped the locations of Alu-derived TTS to their positions in the Alu subfamily consensus sequences [60]. Previously, when a few hundred Alu-TTS were considered as an ensemble, they were found to terminate human gene transcription non-randomly at two specific locations along their sequence [172, 173]. For this study, by considering thousands of Alu-TTS among individual Alu subfamilies of different relative ages, we were able to tease apart this apparently bimodal pattern of termination and discern its origins. The modern Alu element is a dimeric sequence composed of two related precursor sequences: a Free Left Alu Monomer (FLAM) and Free Right Alu Monomer (FRAM) [9, 10]. These sequences themselves descended from the Fossil Alu Monomer (FAM), which in turn descended from a 7SL RNA [10]. Elements from all three families of Alu precursors terminate transcription at single site near their 3'-end (Figure 7.5c). However, when the FLAM and FRAM monomers are considered with respect to their homologous locations in the descendent Alu dimer sequences, these individual termination sites yield a pair a corresponding termination sites; one internal termination site corresponding to the FLAM 3' site and a 3' termination site corresponding to the FRAM 3'site. In modern Alus, the 3' termination site predominates over the internal site and the use of the internal site markedly decreases among younger element sequences (Figure 7.5c). The attenuation in the strength of this TTS donating site from the internal region of the element may reflect the need of the elements themselves to produce full-length transcripts in order to be transposed. In this case, selection against the internal TTS site would be at the level of the element as opposed to

at the level of the host. Thus, the steady migration over time of the Alu-TTS donating site to the 3' end of the element reflects a complex dynamic between inter-element selection and the effects that the elements can in turn exert on their host genome.

Relative levels of utilization for TE-derived TTS

The eight human TE families evaluated here have diverse evolutionary origins, methods of transposition and sequence composition. Given these differences, it would be reasonable to expect that TTS derived from the different TE families would behave differently. To assess whether this is the case, we compared the strength of utilization (see Methods) for TTS derived from members of the different TE families along with the utilization levels seen for non TE-TTS. Individual TTS derived from Alu insertions, while being by far the most abundant TE-derived TTS in the genome (Table 7.1 and Figure 7.4) are utilized far less frequently than TE-TTS derived from other families or non TE-TTS (Figure 7.6). This finding is in accordance with the weak transcription termination previously seen for Alus [178]. On the opposite extreme, TTS derived from sense LTR element insertions, including both the ERV and MaLR families, are utilized significantly more frequently than TTS from any other TE family or alternative non TE-TTS. Indeed, 25% of TTS derived from LTR element insertions have a maximum utilization of over 90% in at least one of the ENCODE cell types. The only group of TTS which shows higher maximum utilization is the group of previously annotated canonical non TE-TTS. The large differences in the relative strengths of Alu and LTR element-derived TTS may explain the differences seen for in the orientation biases between these families (Figure 3). The idea being that Alu insertions provide weaker TTS, and thus may more tolerated in the sense orientation, while ERV and MaLR insertions provide

strong TTS and thus sense oriented LTRs are strongly selected against. Overall, the Alu, ERV and MaLR TE families all exert substantial effects on the expression of human genes via the termination of transcription, but they do so using distinct genome-wide metastrategies. The Alu family, by providing many relatively weak TTS, can affect a large number of genes albeit in a subtle way on a gene-by-gene basis, whereas LTR elements have much larger effects on the expression levels of a smaller number of genes.

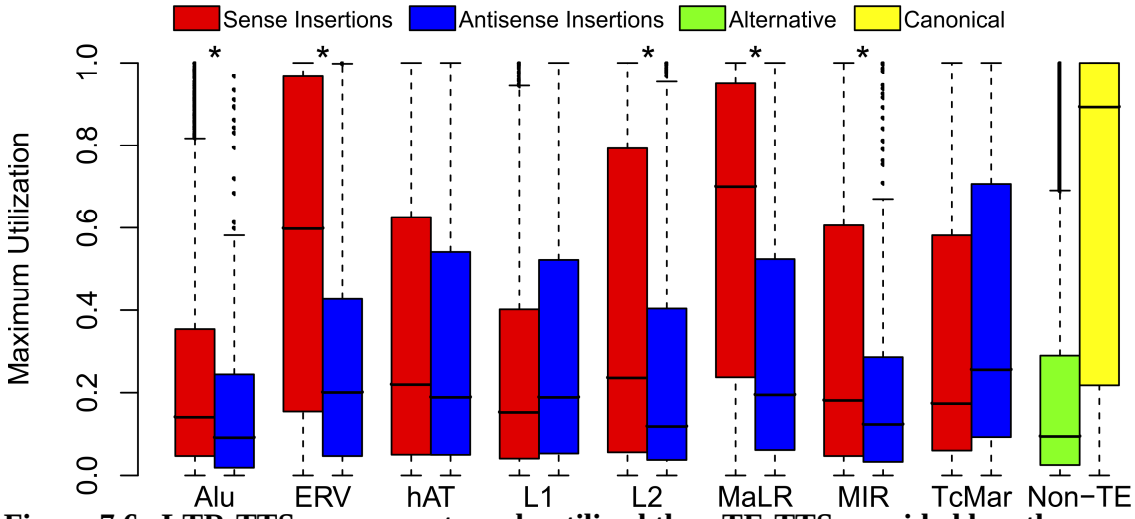


Figure 7.6. LTR-TTS are more strongly utilized than TE-TTS provided by other families. Distributions of maximum utilization values (see Methods) are shown for TE-TTS from different families along with alternative (green) and canonical annotated (yellow) TTS. TE-TTS maximum utilization values are shown separately for sense (red) and antisense (blue) insertions. Statistical significance levels for the differences between the maximum utilization insertion orientations for each TE family (* indicates P<.005) were determined using a Wilcoxon rank-sum test.

The L1 family is curious, being the only TE family to show a strong antisense bias for those insertions providing a TTS (Figure 7.3), yet at the same time showing no difference in TTS strength of utilization between sense and antisense insertions (Figure 7.6). Han *et al.* showed that L1 insertions are capable of terminating transcription in either the sense or antisense orientation, with several polyadenylation signals occurring in the antisense orientation [14]. The same study also showed that sense L1 insertions can cause transcriptional disruption when in the sense orientation, independent of polyadenylation. As the PET technique requires that transcripts be polyadenylated, the data used here cannot take into account non-polyadenylated transcriptional disruption by L1s. Therefore, the anomalous L1 patterns observed here with respect to both TTS

orientation bias and strength of utilization may reflect the relative usage of polyadenylation in L1-TTS from the different strands.

In light the results on the orientation bias of TE-TTS (Figure 7.3), we also compared the strength of utilization for TE-TTS found in sense versus antisense orientations relative to the direction of transcription. Five out of eight of the TE families (Alu, ERV, L2, MaLR and MIR) showed a significant difference ($P< 0.01$, Wilcoxon rank-sum test) in TTS strength of utilization depending on the orientation of the insertion. In all five of these families, TTS derived from sense insertions are more likely to be utilized than those derived from antisense insertions (Figure 7.6). These results are consistent with the findings from the overall TE orientation bias in human genes suggesting that selection acts to remove TE-derived terminators that disrupt gene expression.

Cell type-specific regulatory potential of TE-TTS

Several features of TE-TTS already described in this report raise the possibility that TE-TTS can provide for cell type-specific regulation of gene expression. For example, the individual cases seen in Figure 7.1 clearly demonstrate cell type-specific termination of transcription by TEs. TEs also provide relatively more alternative TTS than non TE-TTS. Finally, individual TE-TTS are utilized less frequently than canonical known TTS from annotated gene models. In order to further investigate the potential genome-wide role of TE-TTS in the cell type-specific termination of transcription, we calculated cell type-specificity levels of all TTS found in genes that are actively transcribed in at least three cell types. The cell type-specificity metric we use measures the extent to which TTS are utilized at different levels across different cell types (see

Methods). TE-TTS show far greater levels of cell type-specificity in the termination of transcription than seen for canonical TTS (Figure 7.7a). In addition, TE-TTS differ in their cell type-specific utilization based on their locations within human genes. Internal TE-TTS, which yield transcripts with truncated ORFs, show significantly more cell type-specific utilization than TE-TTS located in 3' UTRs or downstream of canonical TTS. The relatively highly cell type-specific utilization of internal TE-TTS suggests that they provide a mechanism for dynamic post-transcriptional regulation of human genes via the production of truncated transcripts. TE-TTS within 3'UTRs and downstream of canonical TTS are also generally more cell type-specific than canonical TTS, though to a lesser extent, and these TTS may be functional in producing longer or shorter 3'UTRs. As discussed previously, variation in 3' UTR length provides for yet another level of post-transcriptional regulation [39, 177].

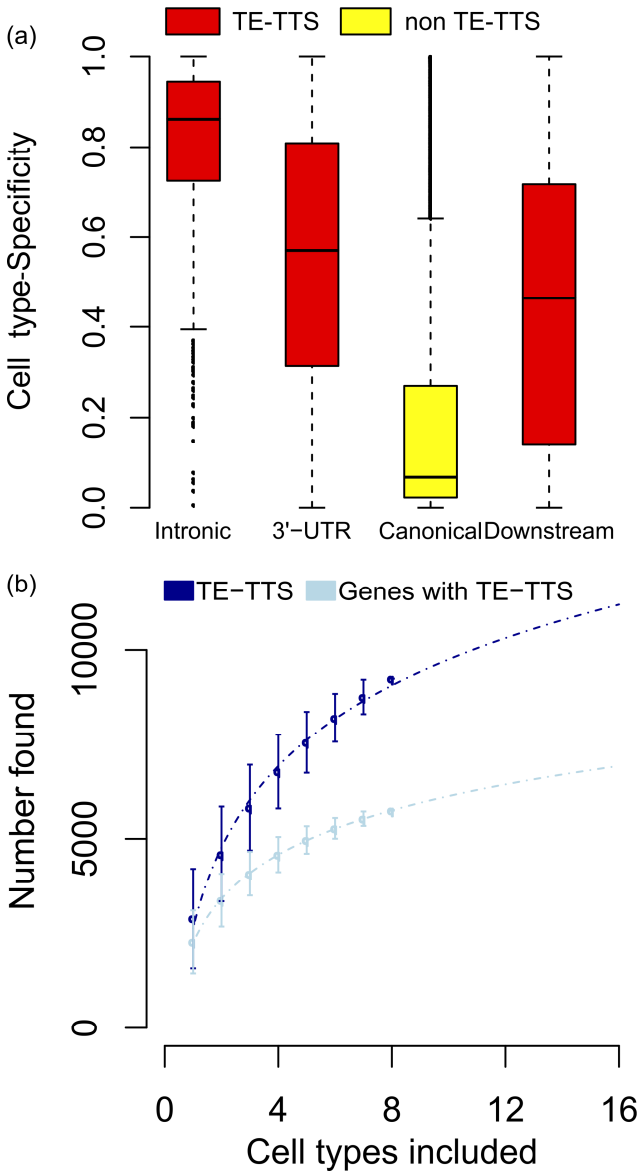


Figure 7.7 TE-TTS terminate transcription in a cell type-specific manor. (a) Cell type-specificity value distributions are shown separately for TE-TTS (red) located within introns, 3' UTRs and downstream of annotated TTS. Cell-type specificity values are also shown for canonical annotated TTS (yellow). (b) Rarefaction curves showing the average numbers (+/- sad) of TE-TTS (dark blue) and genes with at least one TE-TSS (light blue) detected when all possible combinations of 1-8 cell types are used. Observed curves are fitted with a logarithmic trend line.

The apparent cell type-specificity of many TE-TTS suggests the possibility that the TE-TTS discovered in this study via the analysis of eight ENCODE cell types represent only a fraction of the total complement of TE-TTS that exist in the human genome. To address this possibility, we computed a rarefaction curve for TE-TTS by calculating the number of unique TE-TTS found using all possible combinations of 1-8 of the cell types analyzed here (Figure 7.7b). We then fit this rarefaction curve with a logarithmic trend line ($y=31.34\ln x+33.61$; $r=0.99$) to evaluate the extent to which the percent of detected TE-TTS is expected to change with increasing numbers of cell types. Based on the observed trend, we estimated that doubling the number of cell types included in a study of this kind would result in only a 20% increase in the number of TE-TTS found, suggesting a substantially diminishing rate of returns, with respect with respect to the discovery of novel TE-TTS, as more cell types are added. Similarly, the number of genes found to contain a TE-TTS leveled off as more cell types were included. Nevertheless, taking 210 as the total number of human cell types indicates that the TE-TTS discovered here represent half of the total number of human gene TE-TTS. Thus, TEs may provide close to 20,000 TE-TTS for ~11,000 human genes.

Conclusions

Transcription termination as the origin of TE antisense orientation bias

It has been appreciated for some time that TE sequences within the introns of human genes show a strong antisense orientation bias [167]. It was proposed that this bias is due to the propensity of the TE sequences to terminate transcription of host genes when inserted in the sense orientation, resulting in selection against such sense oriented insertions [70]. Nevertheless, studies to date on the ability of TEs to terminate

transcription have not revealed evidence in support of this hypothesis [172, 173]. Here, for the first time, we provide genome-scale evidence in support of the notion that the antisense orientation bias of TEs can be attributed to their ability to preferentially terminate host gene transcription when inserted in the sense orientation. We have shown that TE sequences which provide a TTS are significantly more likely to be found in the sense orientation than other intragenic TE sequences (Figure 7.3), and that TE-TTS in the sense orientation terminate transcription much more efficiently than those found in the antisense orientation (Figure 7.6).

Among the eight TE families studied here, the Alu, ERV and MaLR families are distinct from the other 5 families. TTS from derived from Alu sequences are generally weakly utilized compared to other TE families, while at the same time having a weak antisense orientation bias. The weaker orientation bias of Alu sequences suggests that there is weaker selection against Alu sequences inserted in sense. We suggest that this weaker selection is due to the generally weak utilization of Alu-TTS. Conversely, LTR elements, the ERV and MaLR families, show a very strong antisense bias and a strong utilization; such strong utilization may account for the strong antisense orientation of LTR elements.

Cell type and lineage-specific termination of transcription by TEs

Evidence reported here points to the contribution of TE sequences to the cell type-specific termination of transcription; we have shown that internal TTS derived from TE sequences are significantly more cell type-specific compared to canonical TTS (Figure 7a). In this way, TE sequences have contributed substantially to the generation of cell-type specific patterns of human gene expression via the pre-mature termination of

transcription. In addition to providing for cell type-specific termination of transcription, data reported here indicate that TE sequences are also likely to have contributed substantially to evolutionary lineage-specific transcription termination. Numerous TE insertions can be generated in a short evolutionary time, and accordingly the majority of human TE subfamilies are lineage-specific specific [101]. This means that the regulatory effects that these TEs exert on their host genomes, including termination of transcription as shown here, will also be lineage-specific and account for regulatory differences between evolutionary lineages.

The Alu family, for example, is a relatively young family of TEs, which is confined to the primate evolutionary lineage. The Alu family has been active throughout primate evolution [1], and has likely been altering primate gene expression via TTS donation since the origin of the primate lineage, as can be seen from the results on the more ancient Alu antecedents from the FAM-related subfamilies (Figure 7.5). This process appears to have accelerated, leading to even more species-specific differences in transcription termination, with the amplification of the more modern Alu dimers (Figure 5).

Transcription termination via TE sequences as a common phenomenon

The abundance of TE insertions across eukaryotic lineages suggests that the effect of TE insertions on gene expression via the termination of transcription is not limited to humans [7]. In this study, we characterized the involvement of eight evolutionary diverse families of TEs in the termination of transcription. TEs related to these eight families are present in present in the genomes of many other eukaryotes. For instance, while LTR elements are functionally dead in humans [1], multiple LTR element families are still

highly active in and other species, *e.g.* the Intracisternal A particle (IAP) family of mouse. Indeed, it has been estimated that 10% of mutations in mouse are caused by the novel retrotransposition of an LTR element. As a consequence of this, mice presumably have to contend with a great deal of deleterious transcription termination via novel LTR element insertions. However, these novel insertions also provide the opportunity for innovation in the regulation of gene expression.

CHAPTER 8

CONCLUSIONS

Studies here to date

Through this dissertation, six studies regarding the effect of non-coding sequences, most notably sequences derived from TE insertions, are shown. These studies revealed novel sources of transcription variation in the human genome derived from non-coding sequences. The initial two studies deal with transcription initiation by human TE sequences; the first (Chapter 2) characterized thousands of novel cis-NAT promoters derived from TE sequences and *associated with* human genes, while the second (Chapter 3) characterized endogenous retroviral sequences which function as promoters *for* human genes. In the third study (Chapter 4), techniques for characterizing transcription factor binding sites derived from TE sequences using data generated from the massively high-throughput ChIP-seq technique were reviewed. Such techniques allow for the characterization of lineage-specific transcription factor binding sites derived from TE sequences. The fourth study (Chapter 5) characterized the chromatin environment of cis-NAT promoters associated with human genes, and shows that the cis-NAT promoters bear activating histone modifications in accordance with their activity. Further it is shown that high cis-NAT and gene promoter expression co-occurs far more often than expected, suggesting some form of co-regulation. Also examining the local chromatin environment of non-coding sequences, the fifth study (Chapter 6) reviewed ways in which the epigenetic modification of TE sequences have been shown to effect the expression of nearby host genes. Finally, whereas several of the earlier studies focused

on the promoter activity of non-coding sequences, the sixth study (Chapter 7) characterized the cell-type specific termination of human gene transcription by TE sequences.

In order to replicate and spread within the host genome, TEs must first be transcribed into mRNA. This means that they need their own promoters and TFBS, *i.e.* the internal promoter of the L1 family [13]. It would not be surprising then, given the number of TE sequences in the human genome, if some were TE sequences were able to promote transcription of not just their own sequence, but of neighboring host sequences as well. In the first two studies shown here (Chapter 2, Chapter 3), we have characterized such phenomena. Antisense transcription in the human genome has been appreciated for some time, though the role of these antisense transcripts, remains largely unknown [33, 34, 42]. In Chapter 2 we showed that a substantial fraction of human antisense transcription is initiated from TE-derived promoters. Indeed, we showed that the large majority of human genes have some level of antisense transcription initiated by a TE sequence. In the second study, instances where members of the ERV family provided promoters to human protein coding gene were characterized; over 100 cases where an ERV sequence could be shown to transcribe a human gene were identified. In both of these studies, the vast majority of the sequences involved can no longer transpose, but they have nevertheless altered the human transcriptome via their promoter activity.

The ChIP-seq technique has revolutionized the understanding of DNA-protein interactions, allowing for the genome-wide characterization of transcription factor binding and histone modifications [27]. Where traditional Sanger sequencing produced reads in excess of 700 base-pairs, *e.g.* for EST sequencing, ChIP-seq reads are often

much shorter, due to (1) the technology used and (2) the need for resolution in measuring the DNA-protein interaction. However, a shorter sequence greatly increases the number of locations in a genome to which the sequence could map. Thus, given the highly repetitive nature of the human genome, DNA-protein interactions involving repeated sequences are more difficult to characterize. However, such characterization is important given that TE sequences can spread TFBS through the genome [2, 24, 100]. The methods presented in Chapter 4 allow for the characterization of TFBS derived from TE sequences. Such TFBS could alter the transcriptome of the host in a lineage-specific manor. Similar techniques using data from different cell types could also be used to identify DNA-protein interactions involving TE sequences that appear in a cell type-specific or condition-specific manor.

In chapter 5 the chromatin environment of cis-NAT promoters as it relates the cis-NAT promoter activity was investigated. The study presented in Chapter 2 and other previous studies have identified many cis-NATs for human genes [34, 42], however the regulation of cis-NAT expression remained largely unexplored. Chapter 5 showed that cis-NAT promoters in the human genome bear chromatin modifications similar to those seen on the promoters of human protein coding genes, and that cis-NAT promoter activity is correlated with the level of activating histone modifications, *e.g.* H3K9Ac. The presence of activating epigenetic marks on these cis-NAT promoters strongly suggests that the activity of these cis-NAT promoters is regulated, and such regulation is indicative of function. Indeed it was shown in Chapter 5 that there is an association between highly expressed genes and highly expressed cis-NATs, suggesting that the two may somehow be linked. These cis-NAT promoters have altered the human

transcriptome first through the transcription of cis-NATs, but also possibly through activation of the associated gene.

The transposition of TE sequences within the host organism will likely be neutral at best and very possibly deleterious. It would not be surprising then if host genomes employ mechanisms for reducing transposition. One such mechanism is the deposition of repressive chromatin modifications onto the TE sequences, preventing transcription of the TE and therefore also preventing transposition. While preventing transposition would be a benefit to the host, the silencing of the TE sequence could conceivably also effect the expression of nearby genes via the generation of heterochromatin. In Chapter 6, we reviewed studies which have shown that the deposition of silencing epigenetic modifications on ERV sequences does in fact have such an effect. Additionally, we reviewed instances of epigenetic activation of ERV insertions which lead to the activation of nearby genes.

While the mechanisms of transcription termination and polyadenylation in eukaryotes have been known for some time, the use of cell type-specific transcription termination as a possible layer of gene regulation has not been appreciated for nearly as long. Two recent studies have shown that alternative TTS are utilized in induced pluripotent stem cells and cancer cells, resulting in the generation of transcripts with intact ORFs, but truncated 3'UTRs [39]. Conversely, it was shown that mouse cells undergoing differentiation utilized alternative TTS which resulted in longer 3'UTRs [41]. Such differential 3'UTR generation could affect the cellular lifespan of the mRNA via the loss or gain of miRNA binding sites and nonsense mediated decay, and thus function in the regulation of gene expression [177]. In Chapter 7, we have greatly added the

knowledge of alternative termination of transcription in humans by characterizing alternative and cell type specific TTS derived from TE sequences. While many of these TE-TTS resulted in alternative 3'UTRs, many more were located within introns of human genes, resulting in transcripts missing part of the ORF. These TE-TTS were highly cell type-specific, suggesting that producing truncated transcripts may serve as a common method of gene regulation. Interestingly, it was recently shown that such internal termination sites could lead to production of truncated messages which when translated lacked C-terminal domains. These truncated proteins have been suggested to act as molecular decoys, reducing the availability of ligands for full-length proteins [180]. The work shown here demonstrates that TE sequences have dramatically altered human transcription via the alternative termination of transcription in a cell type-specific manor.

Future prospects of genome-wide bioinformatics studies

The bioinformatics studies presented here yielded interesting results, but the level at which they studied the function human genome was quite high, *i.e.* from a bird's eye view; many potentially interesting non-coding elements, which have many potential effects on human gene transcription, were shown, but the biological effects uncovered computationally were not experimentally demonstrated. The field of bioinformatics should be, at this point, mature enough to move beyond such high-level studies and probe more deeply for specific biological effects and biological meaning. For example, the FLAM_C-derived TTS in the BSDC1 3'UTR (Figure 7.1) is very intriguing and has significant potential to affect the cell type-specific regulation of BSDC1 protein production via 3'UTR shortening and miRNA binding site loss. Further, such effects could happen only in primates where the FLAM_C insertion is found, meaning that any

resulting regulation would occur only in primates. This raises the possibility that this insertion has helped to shape primate-specific gene expression. However, there exists at this point no evidence that 3'UTR shortening at the BSDC1 locus has any such effect on the amount of BSDC1 protein produced. A relatively simple set of experiments could be conducted in order to quantify the affect that the BSDC1 3'UTR has on protein levels, similar to what was done by Jenal et al. [181]; though the study by Jenal et al. had very different goals, a similar set of constructs would be useful.

Here, I provide an example of the kind of experimental studies that could be used to validate the bioinformatic predictions I made in my Ph.D. dissertation research via the analysis of using genome-wide data sets. Several constructs for expressing *Renilla* luciferase with different segments of the BSDC1 3'UTR could characterize the ability of the different segments to lower mRNA and protein levels (Figure 8.1). The polyadenylation signals of the FLAM_C-derived TTS and BSDC1 canonical TTS would be mutated and a very strong polyadenylation signal added to the 3'-end of the construct (black) to ensure that only a singly mRNA form is produced from the construct. The FUTR (FLAM_C-UTR, red) is the truncated 3'UTR that the BSDC1 transcript would contain if the FLAM_C-derived TTS were utilized. The mUTR (miRNA-UTR, yellow) is the portion of the BSDC1 3'-UTR downstream of the FLAM_C-derived TTS containing the miRNA binding sites. Cells would be transfected with these constructs and a *Firefly* luciferase expression construct.

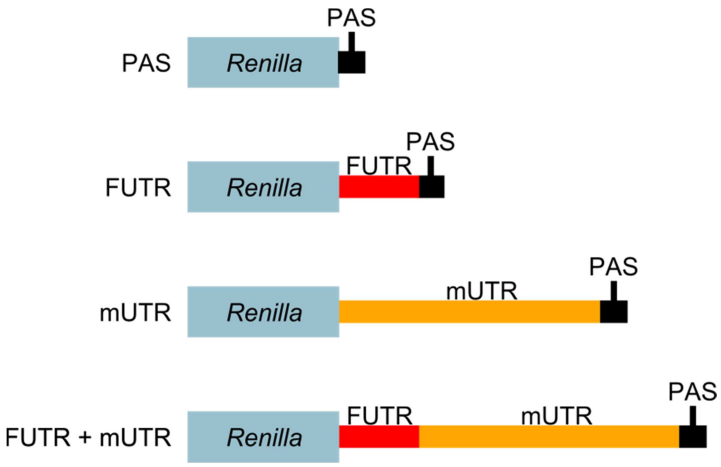


Figure 8.1. Constructs for characterizing the effect of the BSDC1 3'UTR on protein levels. The *Renilla* luciferase gene (blue) is paired with different combinations of the BSDC1 3'UTR: the 3'UTR generated when the FLAM_C-derived TTS is utilized (FUTR, red) and the remainder of the canonical 3'UTR containing miRNA binding sites (mUTR, yellow). All constructs contain a strong polyadenylation signal (black).

Such experiments would be carried out in two cell types: one in which some or all of the miRNAs which have target sites in the BSDC1 3'UTR are expressed, and one in which they are not. One or more of the miRNAs would be artificially expressed in the cell type lacking miRNA expression in order to demonstrate that the effects of 3'UTR shortening on protein levels are a result of miRNA binding. Alternatively, a highly expressed miRNA sponge, in this case the mUTR, could be introduced into the cell type expressing the miRNAs in order to reduce the effective miRNA expression [182, 183]. Hypothetical *Renilla*/*Firefly* luciferase ratios relative to the construct lacking a BSDC1 3'UTR fragment (PAS, Figure 8.1) are shown in Figure 8.2. For all three conditions, it would be expected that the relative *Renilla/Firefly* ratio for the FUTR construct, being short and lacking any known miRNA binding sites, would be close to 1, *i.e.* the same as the PAS construct. In the cell type lacking relevant miRNA expression, the levels relative *Renilla*/*Firefly* ratios for the mUTR and FUTR + mUTR constructs would be

expected to be similar as well, possibly slightly lower due to the unusually long BSDC1 3'UTR. However, for a cell type naturally expressing miRNAs targeting the BSDC1 3'UTR, or for the cell type artificially expressing them, the relative *Renilla*/*Firefly* luciferase ratio would be expected to be much lower. The cell type expression the mUTR miRNA sponge would be expected to show a higher relative *Renilla/Firefly* luciferase the mUTR containing constructs due to the 'sponging' of the relevant miRNAs. A similar analysis via qRT-PCR would need be done to characterize the mRNA levels of *Renilla* and *Firefly* luciferase produced. Similar experiments could be conducted for other genes in which a 3'UTR TE-TTS results in the loss of miRNA binding sites or otherwise dramatically altered 3'UTRs.

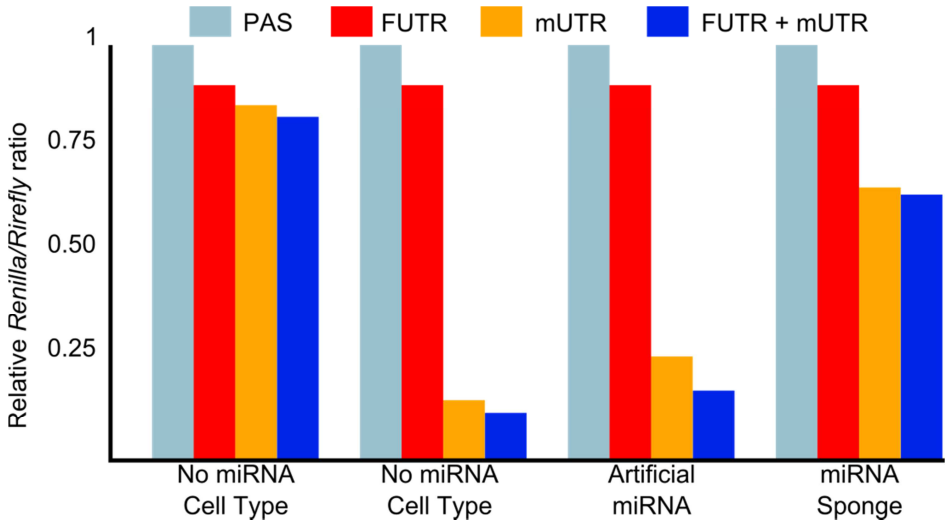


Figure 8.2. Hypothetical relative *Renilla*/*Firefly* luciferase activity levels. The expected relative ratios of *Renilla*/*Firefly* luciferase with the *Renilla* construct containing different sections of the BSDC1 3'UTR.

Such additional experimental investigation may become the norm for bioinformatics studies of functional genomics; the genome-wide analysis of non-coding functional elements, while academically interesting, could be made even more relevant vie the experimental elucidation of specific, verified, effects. Returning to the BSDC1 locus, the differential utilization of the FLAM_C-derived TTS would be of little consequence if it had no overall effect on the expression or translation of the BSDC1 gene; it would be in effect a molecular curio. The need to attach experimentally validated biological meanings to bioinformatics results will likely increase as the field matures and standards for publication rise. Indeed, it may become, or perhaps already has become, the main role of bioinformatics to uncover those previously unnoticed yet interesting functional elements that can later be characterized via classic molecular experimental techniques.

APPENDIX A

SUPPLEMENTARY INFORMATION FO CHAPTER 3

Supplementary Methods

Paired end ditags (PETs)

Gene identification signature (GIS) analysis is a sequencing and mapping strategy that allows for the high-throughput demarcation of gene transcription boundaries, *i.e.* the 5' and 3' gene termini [80]. The GIS analysis procedure that produced the data we analyzed started with the isolation of polyA+ RNA from cells lines subject to different treatments: 1) the log phase of MCF7 cells, 2) MCF7 cells treated with estrogen (10nM beta-estradiol) for 12 hours, 3) HCT116 cells treated with 5FU (5-fluorouracil) for 6 hours, 4) the log phase of embryonic stem cell hES3 in feeder free culture condition. Full-length cDNAs (flcDNA) were generated from RNA and selected using the biotynlated CAP trapper method [184]. The CAP trapper method relies on the introduction of a biotin group to the cap structure found at the 5' end of full length mRNAs followed by first strand cDNA synthesis. Biotin residues are selected using streptavidin-coated magnetic beads, which results in the retention of only flcDNAs. BamHI and MmeI restriction sites are ligated to the 5' and 3' termini of the flcDNAs, which are then cloned to produce the GIS-flcDNA library. This library is digested with MmeI to yielding 18bp sequence fragments (signatures) from the 5' and 3' ends of flcDNAs. The 3' end of the signature includes two A residues from the polyA tail. The

5’ and 3’ flcDNA signatures are covalently ligated to form 36bp paired-end ditags (PETs), each of which represents an individual transcript. PETs are exised using BamHI digestion and then concatenated and cloned for high-throughput sequencing. A single sequencing read of ~700bp leads, on average, to the characterization of 15 distinct PETs.

The GIS cloning and sequence analysis resulted in 584,624 PETs for the log phase MCF7 cells, 153,179 PETs for the estrogen-treated MCF7 cells, 280,340 PETs for the HCT116 cells, and 1,799,970 PETs for the hES3 cells. These PETs were then mapped to the human genome using the following criteria: paired 5’ and 3’ ends must be on the same chromosome, they must be in the correct 5’-to-3’ order and orientation, they must be within 1 million base pairs, there must be a 16bp contiguous sequence match (out of 18bp) for the 5’ end of the PET and a 14bp contiguous match (out of 16bp) for the 3’ end of the PET. Using these criteria, most of the PET sequences (>90%) mapped to single locations in the human genome, but PETs mapping to 2-10 locations were also included in the analysis.

The quality and mapping specificity of PETs has been confirmed in a number of different ways [80]. For instance, >95% of PETs map to known human gene transcripts and the vast majority fell within 10bp of the transcription start and termination sites. Most relevant to our study is the fact that the GIS analysis has been shown to be 30 times more efficient than standard cDNA methods for characterizing transcript and has resulted in the discovery of numerous previously uncharacterized transcripts. Thus, GIS is particularly suited to the discovery of alternative transcripts in the human genome of the kind initiated by ERV sequences.

Cap Analysis of Gene Expression (CAGE)

The CAGE technique was developed for the high-throughput characterization of transcription start sites (TSS) [30]. CAGE uses a similar technology to that described above for the generation of PETs in GIS. The main difference is that CAGE only characterizes the 5' ends, as opposed to both 5' and 3' PET ends, of flcDNAs. CAGE also employs the isolation of flcDNAs using biotinylated mRNA caps as described for GIS. Once flcDNAs are isolated, linkers with MmeI restriction sites are ligated to the 5' ends of the flcDNAs, and the first 20 bp of the cDNAs is cleaved with a MmeI restriction digest. The resulting 5' end cDNA fragments (so-called CAGE tags) are amplified, concatenated and sequenced. This procedure allows for the high-throughput characterization of the 5' ends of mRNAs, and mapping of the resulting sequence fragments to the genome identifies transcriptional start sites (TSS). CAGE tags are mapped to the human genome mandating a contiguous match of 18 out of 20bp. Approximately 60% of CAGE tags can be unambiguously mapped to the genome in this way. Only CAGE tags that mapped to one location in the genome were used in our study.

CAGE is a slightly more mature technology than GIS and it has been extensively validated [31, 54]. In addition to the ability of CAGE tags to converge on known TSS in the human genome, CAGE also identifies thousands of previously unknown TSS. This is consistent with our discovery that numerous ERV-derived TSS correspond to alternative transcripts.

Gene expression analysis

Human and mouse gene expression data were taken from the Novartis mammalian gene expression atlas version 2 (GNF2) [87]. GNF2 data are based on

Affymetrix microarray experiments conducted in replicate on 79 human and 61 mouse tissues. For each Affymetrix probe, signal intensity values (*i.e.* expression levels) were median and log2 normalized across tissues. Affymetrix probes were mapped to GenBank RefSeq gene accessions using the UCSC Table Browser utility [81]. Human-mouse orthologous gene pairs and 28 corresponding tissue pairs were identified as described previously [185]. Similarity between human-mouse orthologous gene pair tissue-specific expression profiles was measured using the Pearson correlation co-efficient (r) as described previously [186]. An adjusted r-value threshold of 0.5789, above which human-mouse orthologous gene pairs can be considered to have correlated expression patterns across n=28 tissues, was computed using the formula $t=r*\text{sqrt}((n-2)/(1-r^2))$, where t follows the Student-t distribution with n-2 degrees of freedom. The r-value threshold was based on a P-value of 0.00125 computed using a Bonferroni correction with the number of comparisons (40) performed (*i.e.* P=0.05/40).

The GNF2 data were also used to compare the values of a number of gene expression parameters for human genes that have ERV-TSS that yield chimeric transcripts (ERV+) versus all other human genes (ERV-) with Novartis expression data. Average values for the following gene expression parameters across the two sets were compared: 1) average expression, 2) maximum expression, 3) breadth of expression and 4) tissue-specificity of expression. Average, maximum and breadth of expression were computed as described previously in [186]. Tissue-specificity was computed using the τ parameter described in [187]. The values of τ range between 0 and 1 with more tissue-specific genes having higher values. Human gene tissue-specific expression profiles from the GNF2 data were used to group genes into 20 clusters of co-expressed genes with

K-means clustering using the program Genesis [188]. The observed counts of ERV+ genes in each of these clusters were compared to the expected counts based on the whole genome distribution using a chi-square test.

Human ESTs were mapped to ERV-derived TSS and associated genes and the tissues (or cell lines) from which they were characterized were determined using the Human ESTs track of the UCSC Genome Browser [55]. The distribution of EST tissue types across alternative versus primary promoters was compared using a joint chi-square test. Observed EST tissues counts for the alternative versus the primary TSS were compared with expected counts based on the pooled tissue counts to compute a chi-square value for each promoter and the joint chi-square probability for the two promoters was computed.

Gene ontology (GO) analysis

The set of human genes with ERV-TSS that yield chimeric ERV-gene transcripts (ERV+) were evaluated for enrichment of biological process and molecular function GO terms using the program using the program GOTree Machine (GOTM) [189]. The GOTM program was used to implement a hypergeometric test comparing GO term frequencies in the ERV+ human gene set against a background set made up of all human genes with corresponding Affymetrix probes. GOTM produces a list of enriched GO terms along with a view of the GO directed acyclic graph (DAG) showing the parent-child relationships among enriched GO terms.

ERV age analysis

ERVs accumulate mutations after inserting into the genome. Thus, the relative ages of ERVs, *i.e.* the time since insertion, can be estimated using the sequence

divergence levels between ERVs and their consensus sequences [1]. ERV-to-consensus divergence levels were taken from the RepeatMasker output. Average levels of ERV-to-consensus divergence were compared for all ERVs, ERVs that overlap with ESTs, ERVs that overlap with CAGE tags and ERVs that overlap with PETs.

Table A.1. List of ERVs that initiate ERV-gene chimerical transcripts along with their associated genes.

	Name	Chromosome	Start	Stop	Gene Accession	Chromosome	Start	Stop
Upstream	LTR41	chr1	17954613	17954613	NM_030812	chr1	17954430	18026143
	MER4A	chr10	106004209	106004209	NM_004832	chr10	106004667	106017199
	LTR12D	chr14	23175701	23175701	NM_005794	chr14	23175421	23184686
	LTR12D	chr14	23175701	23175701	NM_182908	chr14	23175421	23184686
	MER54B	chr16	751307	751307	NM_005823	chr16	751133	758866
	MER54B	chr16	751307	751307	NM_013404	chr16	751133	758866
	MER39	chr17	7392869	7392869	NM_003809	chr17	7393098	7401930
	MER39	chr17	7392869	7392869	NM_172089	chr17	7393139	7405649
	MLT2E	chr19	9112101	9112101	NM_020933	chr19	9112072	9135082
	LTR12B	chr3	129354764	129354764	NM_021937	chr3	129355002	129610178
	LTR54	chr4	84425703	84425703	NM_015697	chr4	84404001	84424988
	MER41D-int	chr4	191144734	191144734	NM_020040	chr4	191140672	191143018
	MER51A	chr7	10979661	10979661	NM_014660	chr7	10980040	11109807
	MER51A	chr7	10979661	10979661	NM_001007157	chr7	10980040	11175766
	LTR43	chr8	143778578	143778578	NM_017527	chr8	143778532	143782611
	LTR75_1	chr8	144192262	144192262	NM_173687	chr8	144192053	144207095
In 5'-UTR	LTR41	chr1	17954234	17954613	NM_030812	chr1	17954430	18026143
	LTR41	chr1	17958237	17958696	NM_030812	chr1	17954430	18026143
	HERVH48	chr1	181939402	181944093	NM_015149	chr1	181871830	182164288
	HERV4_I	chr11	117598355	117598585	NM_001098526	chr11	117569651	117601019
	HERVL40	chr12	31766377	31766997	NM_207337	chr12	31715338	31773251
	LTR40a	chr13	29764513	29764758	NM_001014380	chr13	29674767	29779163
	LTR40a	chr13	29764513	29764758	NM_032116	chr13	29674767	29779584
	LTR12D	chr14	23176754	23177445	NM_005794	chr14	23175421	23184686
	LTR12D	chr14	23176754	23177445	NM_182908	chr14	23175421	23184686
	MLT2E	chr19	9111991	9112101	NM_020933	chr19	9112072	9135082
	MER67D	chr19	42033104	42033171	NM_003419	chr19	42033106	42062310
	MER52D	chr2	109729279	109729386	NM_023016	chr2	109729199	109733852
	Harlequin	chr2	188083923	188088834	NM_001032281	chr2	188051551	188127464
	LTR16C	chr20	4666836	4667196	NM_177549	chr20	4659928	4669314
	LTR7	chr4	89292502	89292904	NM_004827	chr4	89230440	89299035
	LTR1D	chr4	178965922	178966704	NM_001085490	chr4	178886900	179148663
	HERV9	chr5	146361399	146364490	NM_181674	chr5	145949260	146415783
	HERV9	chr5	146361399	146364490	NM_181678	chr5	145949260	146441207
	LTR5B	chr5	177414929	177415182	NM_001080544	chr5	177414995	177415888
	LTR50	chr8	12873949	12874595	NM_020844	chr8	12847553	12931655
	LTR43	chr8	143778243	143778578	NM_017527	chr8	143778532	143782611
	HERVH	chrX	113744197	113747503	NM_000868	chrX	113724806	114050880

	LTR41	chrX	134693953	134694178	NM_152582	chrX	134693879	134701914
	LTR41	chrX	134693953	134694178	NM_001017436	chrX	134693879	134719184
	LTR41	chrX	134781361	134781586	NM_001007551	chrX	134773630	134781660
	LTR41	chrX	134798611	134798836	NM_001017438	chrX	134790881	134798910
In CDS	MER54B	chr16	725416	725902	NM_022493	chr16	719771	730998
	LTR9	chrX	2729224	2729740	NM_175569	chrX	2680114	2743960
	MER51A	chr2	3339561	3340106	NM_003310	chr2	3171749	3360605
	MER21C	chr7	5933542	5933818	NM_001099697	chr7	5932302	5976840
	MER21C	chr7	5933542	5933818	NM_173565	chr7	5932302	5976840
	MER31B	chr8	11702728	11703173	NM_004462	chr8	11697598	11734226
	LTR36	chr22	17028682	17028800	NM_017414	chr22	17012757	17040162
	MER41A-int	chr12	26797075	26797151	NM_002223	chr12	26379553	26877398
	LTR12	chr12	29480709	29481473	NM_183378	chr12	29471755	29541886
	HERVE	chr12	31165336	31165939	NM_001080502	chr12	31158726	31250355
	LTR27B	chr7	33358011	33358250	NM_001033604	chr7	33135676	33612205
	LTR27B	chr7	33358011	33358250	NM_001033605	chr7	33135676	33612205
	LTR27B	chr7	33358011	33358250	NM_014451	chr7	33135676	33612205
	LTR27B	chr7	33358011	33358250	NM_198428	chr7	33135676	33612205
	MLT2B2	chr17	34767819	34768179	NM_032875	chr17	34670366	34811402
	MER41G	chr22	34984494	34985070	NM_003661	chr22	34979069	34993522
	MER41G	chr22	34984494	34985070	NM_145343	chr22	34979069	34993522
	LTR7	chr18	38577764	38578263	NM_002930	chr18	38577189	38949655
	MER68	chr4	38951204	38951750	NM_025132	chr4	38860418	38963824
	LTR19C	chr13	42526886	42527709	NM_013238	chr13	42495361	42581304
	MER61F-int	chr15	43131729	43132168	NM_003104	chr15	43102632	43154331
	MER92C	chr4	46920855	46921054	NM_000812	chr4	46728335	47123202
	LTR12C	chr13	50224579	50226004	NM_198989	chr13	50184759	50315886
	MER4C-int	chr7	50525903	50526498	NM_000790	chr7	50493627	50596262
	MER4C-int	chr7	50525903	50526498	NM_001082971	chr7	50493627	50600648
	MER4D	chr3	54650117	54651004	NM_018398	chr3	54131732	55083622
	MER21C	chr5	54816203	54816867	NM_003711	chr5	54756441	54866630
	MER21C	chr5	54816203	54816867	NM_176895	chr5	54756441	54866630
	MER34B	chr4	62321975	62322273	NM_015236	chr4	62045433	62620762
	HERV4_I	chr19	63452392	63453769	NM_014480	chr19	63431881	63466820
	MER57A-int	chr6	64070948	64072854	NM_016571	chr6	64047518	64087841
	MER52A	chr4	64888385	64889631	NM_001010874	chr4	64826015	64957773
	LTR12	chr7	68888991	68889595	NM_015570	chr7	68702254	69895790
	MER52A	chr13	69395827	69397313	NM_020866	chr13	69172726	69580460
	MER4B-int	chr12	74045504	74047192	NM_152779	chr12	74014729	74050436
	MER34B	chr9	74532752	74533337	NM_138691	chr9	74326536	74641087
	MLT2F	chr7	75145897	75146022	NM_005338	chr7	75001344	75206215

	MLT2B2	chr12	79840894	79841339	NM_004664	chr12	79715301	79855825
	HERV17	chr8	81814505	81817335	NM_001033723	chr8	81713323	81949571
	LTR54	chr1	85637738	85638265	NM_012137	chr1	85556756	85703411
	HERVH	chr10	92557476	92561145	NM_000872	chr10	92490557	92607651
	HERVH	chr10	92557476	92561145	NM_019859	chr10	92490557	92607651
	HERVH	chr10	92557476	92561145	NM_019860	chr10	92490557	92607651
	HERVH	chr4	93581454	93584375	NM_001510	chr4	93444572	94912672
	LTR9	chr7	98858058	98858575	NM_015545	chr7	98854689	98874355
	MER34D	chr13	99168644	99168812	NM_206808	chr13	99056936	99342824
	HERVH	chr14	101779943	101781050	NM_014226	chr14	101764930	101841284
	MLT2D	chr7	110096425	110096839	NM_032549	chr7	110090345	110989583
	LTR12C	chr4	110124142	110125521	NM_198721	chr4	109954420	110443248
	LTR12C	chr4	110124142	110125521	NM_032518	chr4	109964489	110443248
	HERVH	chr8	110382902	110385440	NM_032869	chr8	110322324	110415491
	LTR16C	chr12	116670959	116671316	NM_173598	chr12	116375221	116777724
	LTR22B	chr10	117136681	117136898	NM_207303	chr10	116843113	117698484
	LTR7Y	chr3	117306893	117307327	NM_002338	chr3	117011839	117647068
	HERVH	chr3	117309183	117312335	NM_002338	chr3	117011839	117647068
	LTR7Y	chr3	117312335	117312765	NM_002338	chr3	117011839	117647068
	LTR16B	chr9	118368734	118369128	NM_198188	chr9	118227327	118489334
	LTR16B	chr9	118368734	118369128	NM_014010	chr9	118227327	119217138
	LTR16B	chr9	118368734	118369128	NM_198186	chr9	118227327	119217138
	LTR16B	chr9	118368734	118369128	NM_198187	chr9	118227327	119217138
	HUERS-P3	chr6	119008185	119016806	NM_001042475	chr6	118892931	119079713
	HUERS-P3	chr6	119008185	119016806	NM_206921	chr6	118919289	119079713
	MER41B	chr8	119012088	119012717	NM_000127	chr8	118880782	119193239
	MLT2B4	chr8	119444631	119445182	NM_001101676	chr8	119270875	119703365
	LTR38B	chr6	119666263	119666850	NM_005907	chr6	119540967	119712625
	LTR12C	chr3	120273950	120275501	NM_152538	chr3	120102170	120347588
	LTR7	chr6	123945179	123945578	NM_006073	chr6	123579181	123999641
	MER52A	chr4	124247306	124248778	NM_145207	chr4	124063674	124460054
	MER21A	chr6	124675036	124675939	NM_001040214	chr6	124166767	125188483
	LTR22C	chr7	126086080	126086529	NM_000845	chr7	125865892	126670546
	LTR40a	chr8	126207807	126207882	NM_173685	chr8	126173276	126448543
	MLT2A2	chr3	126756469	126756970	NM_022776	chr3	126730393	126796624
	LTR10C	chr5	133976397	133976985	NM_001033503	chr5	133970018	133996426
	LTR10C	chr5	133976397	133976985	NM_016103	chr5	133970018	133996426
	PABL_A	chr9	135137645	135138270	NM_020469	chr9	135120383	135140451
	MER21C	chr2	137646118	137646928	NM_001080427	chr2	137464931	138151757
	HERV9	chr5	146361399	146364490	NM_181676	chr5	145949260	146415671
	HERV9	chr5	146361399	146364490	NM_181677	chr5	145949260	146441207

	LTR1B	chr7	146429315	146430028	NM_014141	chr7	145444385	147749019
	LTR8A	chr7	147352928	147353621	NM_014141	chr7	145444385	147749019
	MER21A-int	chr4	147875445	147878401	NM_031956	chr4	147847628	148086484
	MER4A1-int	chr3	151891622	151892158	NM_152394	chr3	151860366	151904432
	HERV30	chr3	155568283	155571737	NM_001038705	chr3	155538155	155630198
	MER41B	chr6	160579903	160580541	NM_003058	chr6	160557783	160599949
	HUERS-P2	chr3	168448261	168451231	NM_024687	chr3	168440778	168580765
	LTR10G	chr3	168474996	168475504	NM_024687	chr3	168440778	168580765
	LTR7	chr4	187400103	187400470	NM_000892	chr4	187385665	187416618
	HERVH	chr4	187402139	187405364	NM_000892	chr4	187385665	187416618
	HERVL40	chr2	202090151	202090443	NM_152525	chr2	202060401	202192146
	MER21C	chr1	223825852	223826508	NM_001008493	chr1	223741156	223907468
	MER21C	chr1	223825852	223826508	NM_018212	chr1	223741156	223907468
	LTR49-int	chr2	231087179	231087883	NM_003113	chr2	230989114	231089486
	LTR49-int	chr2	231087179	231087883	NM_001080391	chr2	230989114	231118561

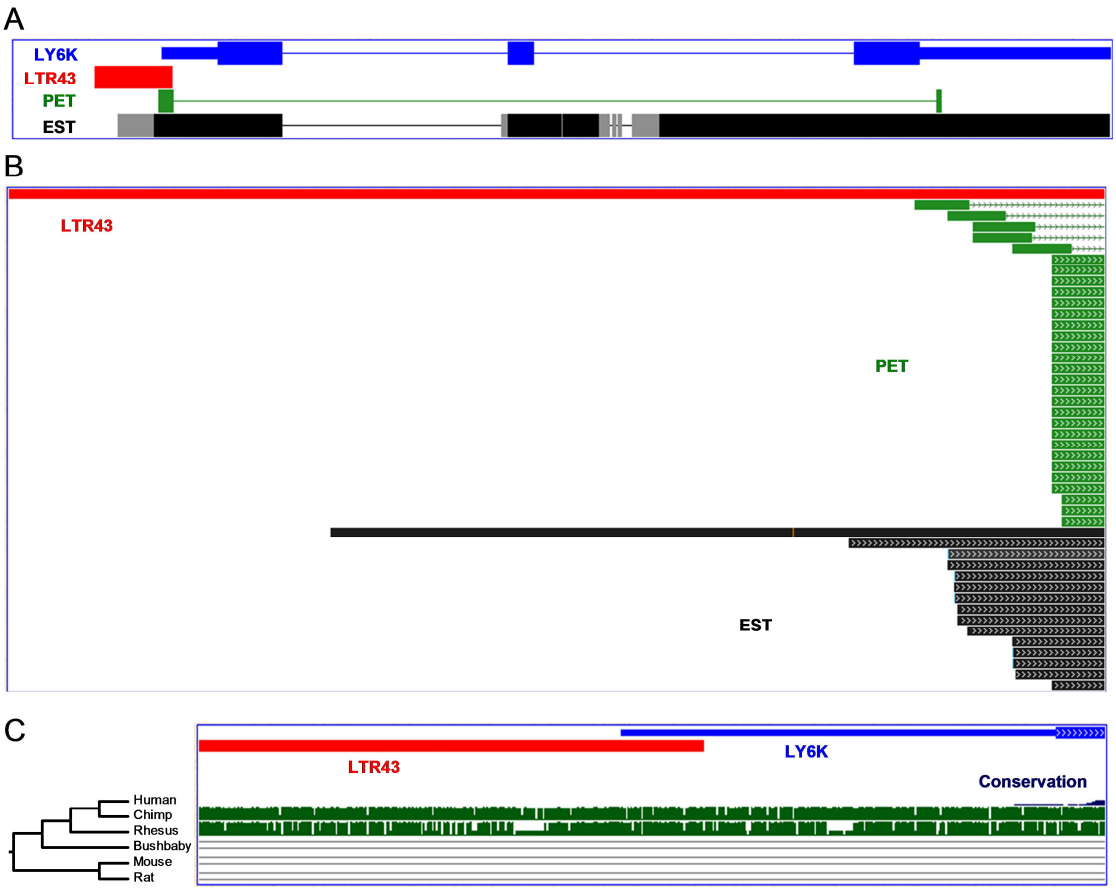


Figure A.1. ERV-derived promoter of the LY6K gene. A) The LTR43 (red) ERV sequence is located in the proximal promoter region and overlaps the LY6K 5' UTR. The locations of PET sequences (green) and spliced ESTs (black) are shown. B) The LTR43 (red) sequence region is enlarged and the individual PET sequences (green) and spliced ESTs (black) that support the existence of this promoter are shown. C) Evolutionary conservation of LY6K versus LY6K.

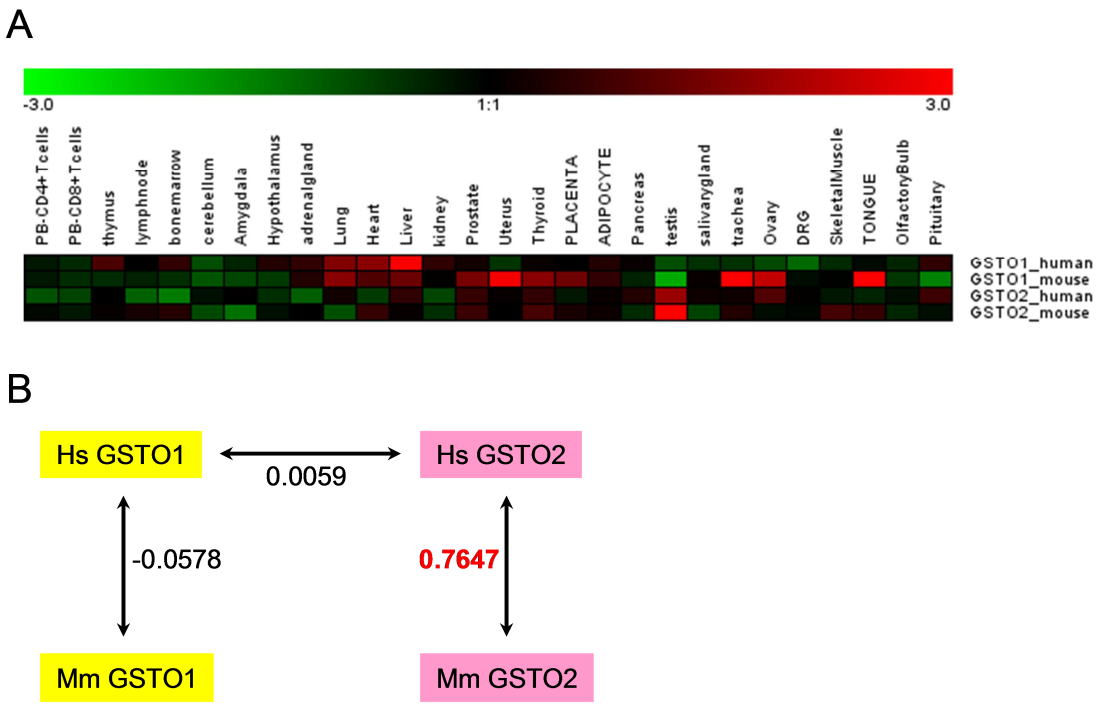


Figure A.2. Gene expression profiles and correlations for human and mouse GSTO1 and GSTO2. A) Relative expression values resulting from median and log2 normalization of Affymetrix signal intensity values across tissues. B) Pearson correlation coefficient values (*r*) showing the correlation, or lack thereof, for tissue-specific expression between human paralogs and human-mouse orthologs.

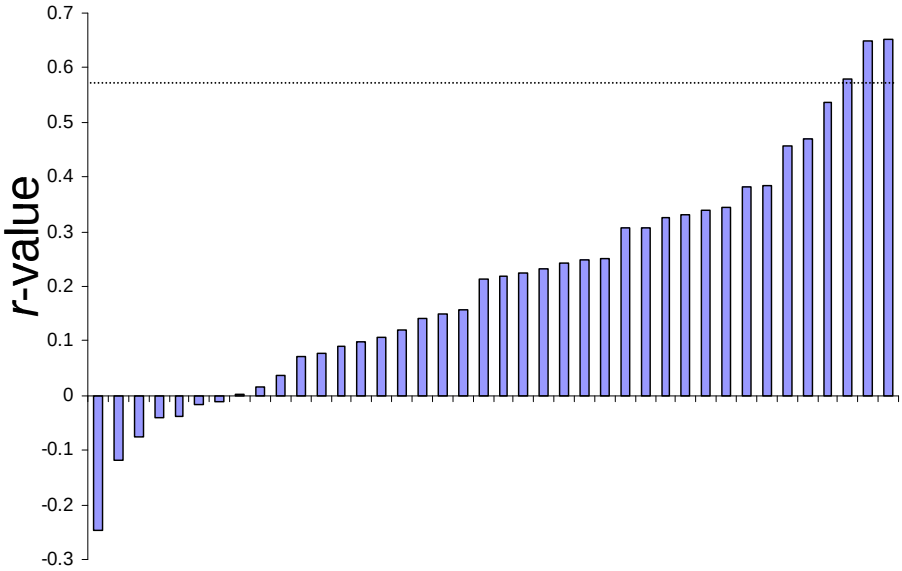
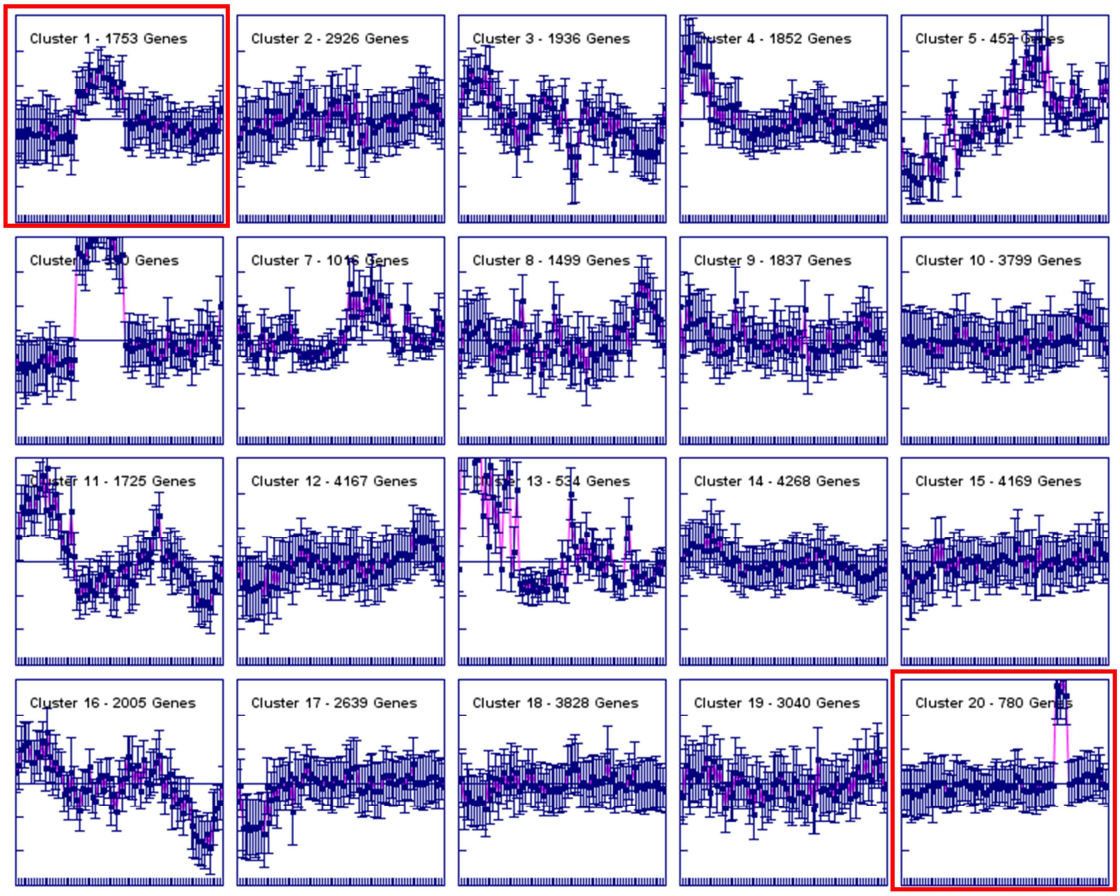


Figure A.3. Ranked list of r-values showing the correlation between human-mouse orthologous gene tissue-specific expression profiles for all human genes that have a lineage-specific ERV-derived TSS that generates a chimeric ERV-gene transcript. An *r*-value≥0.5789, dotted line, corresponds to significantly co-expressed orthologous gene pairs.

Table A.2. Human gene expression values for genes with ERV-TSS versus all other genes.

Expression[a]	ERV+[b]	ERV-[c]	*T*[d]	*P*[d]
Average	378.3 ± 52.9	600.6 ± 10.0	3.97	7.3e-5
Maximum	1920.1 ± 309.9	3143.5 ± 50.33	3.76	1.7e-4
Breadth	23.9 ± 2.6	27.0 ± 0.1	1.18	2.4e-1
Tissue-specificity	0.75 ± 0.01	0.71 ± 0.00	2.88	4.0e-3

[a]Expression parameters measured using the Novartis GNF2 data as described
[b]Average and standard error for human genes possessing an ERV that promotes an ERV-gene chimeric transcript
[b]Average and standard error for all other human genes
[d]Test statistic and significance level for the Student's *t*-test comparing the ERV+ and ERV- values



cluster	#probe	Observed	Expected	chi-square	
1	1753	12	3.87598	17.02787535	brain
2	2926	4	6.469548	0.9426727	
3	1936	1	4.280603	2.514214961	
4	1852	4	4.094874	0.00219815	
5	452	2	0.999397	1.001810501	
6	550	2	1.21608	0.5053366	
7	1016	4	2.246432	1.368837315	
8	1499	1	3.314372	1.616088155	
9	1837	3	4.061709	0.277524844	
10	3799	6	8.399799	0.685615837	
11	1725	1	3.81407	2.07625744	
12	4167	4	9.213467	2.950055682	
13	534	1	1.180704	0.02765619	
14	4268	7	9.436784	0.629230882	
15	4169	14	9.217889	2.480891257	
16	2005	7	4.433166	1.486215024	
17	2639	7	5.834975	0.23261172	
18	3828	5	8.46392	1.417633857	
19	3040	6	6.721608	0.077469284	
20	780	8	1.724623	22.83418023	testis
	44775	99		60.15437598	Chi-square value
				3.65815E-06	*P*-value

Figure A.4. Co-expressed clusters of human genes. Average tissue-specific expression profiles across 79 tissues are shown for each cluster. Clusters enriched for genes with ERV-TSS that generate chimeric transcripts are boxed in red. Chi-square statistical analysis indicating enrichment in cluster 1 (brain) and cluster 20 (testis) is shown below the clusters.

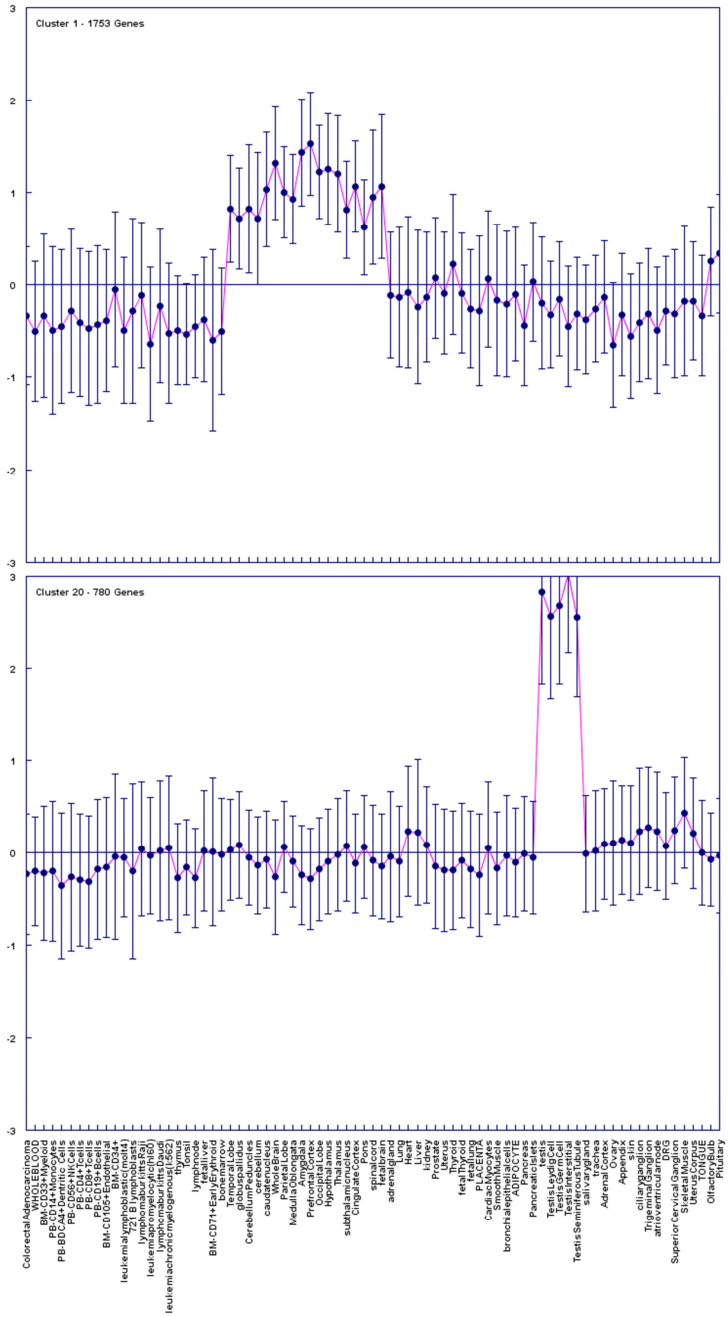


Figure A.5. Human gene co-expression cluster 1 (brain) and cluster 20 (testis) are shown. Average relative expression levels are indicated on the y-axis and the tissue-names are shown on the x-axis below the second panel.

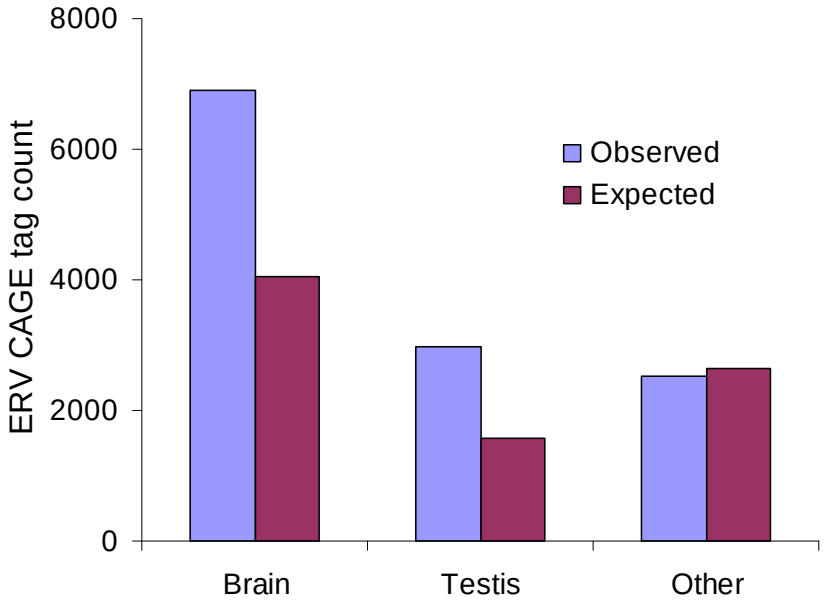


Figure A.6. Tissue distribution of ERV CAGE tags. Observed counts for ERV CAGE tags are compared to expected counts based on all CAGE tags for brain, testis and the average of all other tissues. χ^2=3,249 *P*=0.

Table A.3. Statistically over-represented (enriched) GO biological process terms for human genes with an ERV-derived TSS generating a chimeric ERV-gene transcript.

AffyID[a]	ERV-gene[b]	GO[c]	*P*-value[d]
201563_at	NM_003104	GO:0019751 polyol metabolic process	0.0015
205311_at	NM_000790, NM_001082971	GO:0006066 alcohol metabolic process	0.0034
206463_s_at	NM_005794, NM_182908	GO:0008202 steroid metabolic process	0.0030
208647_at	NM_004462	GO:0008299 isoprenoid biosynthetic process	0.0013
209546_s_at	NM_003661, NM_145343	GO:0008202 steroid metabolic process	0.0030
210946_at	NM_003711, NM_176895	GO:0044255 cellular lipid metabolic process	0.0089
213379_at	NM_015697	GO:0008299 isoprenoid biosynthetic process	0.0013
218304_s_at	NM_022776	GO:0008202 steroid metabolic process	0.0030

[a]AffyID mapped to ERV-related gene
[b]ERV-related gene
[c]Over-represented biological process GO term and description
[d]*P*-value associated with that GO term

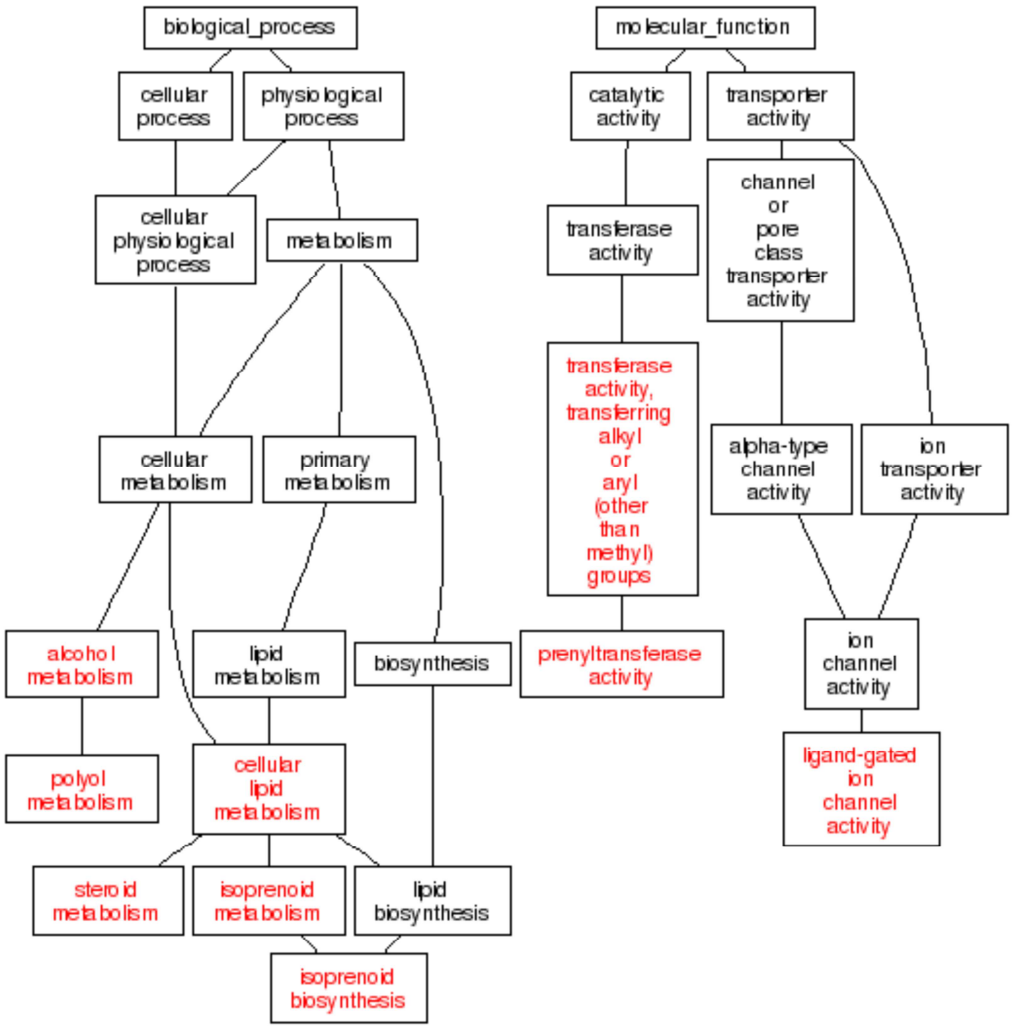


Figure A.7. GO directed acyclic graph showing the parent-child relationships of statistically over-represented (enriched) GO biological process and molecular function terms for human genes with an ERV-derived TSS generating a chimeric ERV-gene transcript. Significantly enriched GO terms are shown in red.

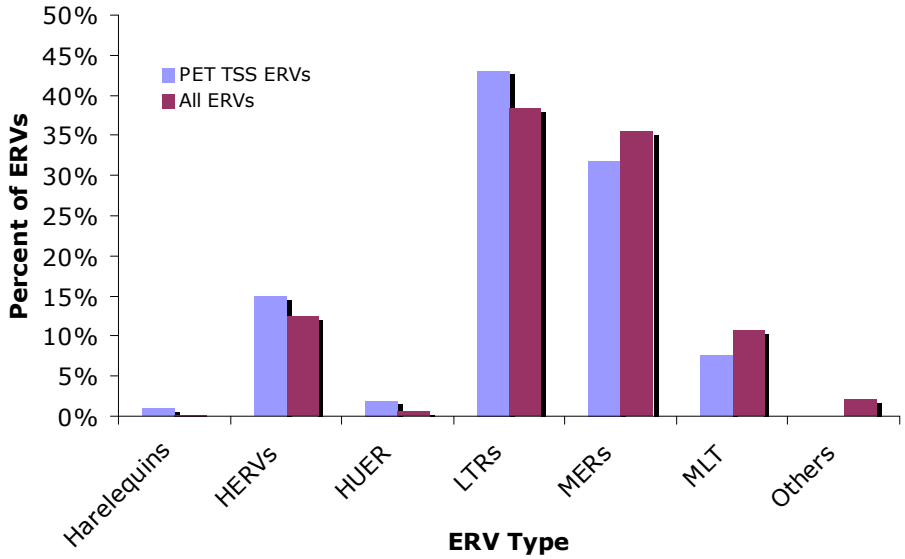


Figure A.8. Relative frequency of ERV-derived TSS detected by PET versus all ERVs in the genome. ERV types correspond to family names from the RepeatMasker output.

09

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table B.1. Number of CAGE tags mapped from each cell line, sub-cellular location and poly-adenylation state. CAGE tag mappings from the GM12878, H1HESC, HEPG2, HUVEC and K562 and NHEK cell lines were downloaded from the ENCODE repository on the UCSC genome browser.

Cell Line	Sub-cellular location	Poly-A-	Poly-A+	Total
GM12878	Cytosolic	18,211,686	---	---
	Nucleolar	---	---	26,792,181
	Nuclear	27,652,635	---	---
H1HESC	Whole Cell	28,801,912	---	---
HEPG2	Cytosolic	19,645,027	---	---
	Nucleolar	---	---	35,803,226
	Nuclear	16,792,966	---	---
HUVEC	Cytosolic	19,837,471	---	---
K562	Cytosolic	20,273,886	18,769,778	---
	Nucleolar	---	---	9,527,032
	Nucleoplasmic	---	---	14,826,128
NHEK	Nuclear	25,989,950	20,648,810	---
	Cytosolic	23,312,041	---	---
	Nuclear	68,757,727	---	---

Table B.2. CAGE clusters identified in each cell line, sub-cellular location and poly-adenylation state. Overlapping CAGE tag mappings (Table S1) from the ENCODE cell lines and in the same sub-cellular locations and poly-adenylation states were grouped together and designated as clusters

Cell Line	Sub-cellular location	Poly-A-	Poly-A+	Total
GM12878	Cytosolic	407,021	---	---
	Nucleolar	---	---	2,458,566
	Nuclear	1,087,671	---	---
H1HESC	Whole Cell	903,838	---	---
HEPG2	Cytosolic	668,040	---	---
	Nucleolar	---	---	2,888,807
	Nuclear	4,188,848	---	---
HUVEC	Cytosolic	857,093	---	---
K562	Cytosolic	4,096,071	525,177	---
	Nucleolar	---	---	3,503,588
	Nucleoplasmic	---	---	4,617,119
NHEK	Nuclear	6,829,025	2,244,742	---
	Cytosolic	1,730,893	---	---
	Nuclear	3,082,557	---	---

Table B.3. ChIP-seq reads mapped for each histone modification and cell types. ChIP-seq data from the GM12878, H1HESC, HEPG2, HUVEC, K562 and NHEK cell types cell ties were downloaded from the ENCODE repository on the UCSC genome browser. Reads were mapped using bowtie, keeping the best hits with ties broken by quality. Ambiguously mapped reads were resolved using GibbsAM.

Modification	Tags Mapped					
	GM12878	**H1HESC**	**HEPG2**	**HUVEC**	**K562**	**NHEK**
Control	7,436,431	11,908,617	11,039,784	16,836,245	13,240,739	10,666,985
H3K4Me1	14,069,086	9,713,507	---	14,524,897	---	11,260,426
H3K4Me2	9,163,434	14,479,372	17,293,347	12,005,596	12,454,360	11,031,009
H3K4me3	10,218,953	7,072,374	10,289,145	12,497,262	15,989,323	10,296,574
H3K9Ac	12,022,891	16,477,468	7,351,567	8,670,429	17,281,199	12,454,536
H3K9Me1	---	---	---	10,658,052	15,905,405	10,731,385
H3K27Ac	10,770,731	---	8,856,877	16,833,005	15,871,535	12,788,055
H3K27Me3	14,430,662	7,160,479	---	11,652,289	12,412,831	9,141,036
H3K36Me3	15,195,406	14,680,520	13,579,529	9,818,236	14,950,529	9,182,104
H4K20Me1	12,224,195	16,605,685	10,356,633	16,664,745	13,685,630	12,380,840
Pol2b	---	---	---	9,860,160	10,822,295	10,175,792

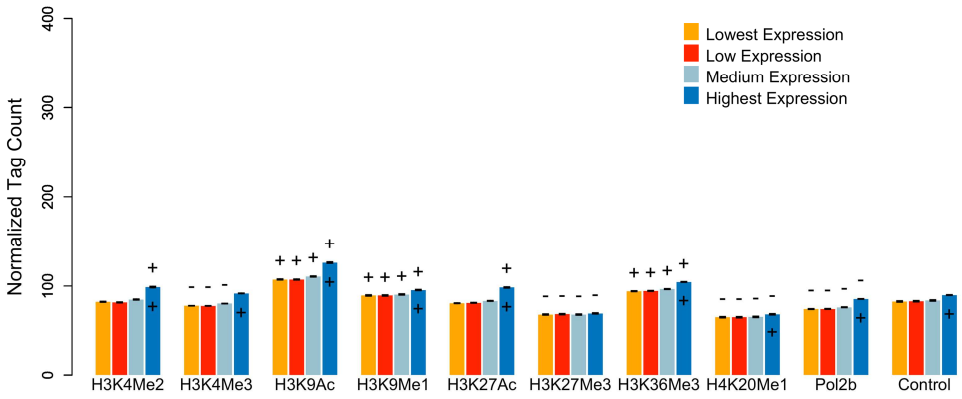


Figure B.1. Enrichment of chromatin modifications and RNA PolII at cis-NAT promoters in K562 cells using CAGE data from nucleus polyadenylated isolates Cis-NAT promoters were divided into 4 bins based on activity, and the normalized average numbers of ChIP-seq reads from each histone modification +/-5kb of the cis-NAT TSS were calculated for each bin. A '+' or '–' above a bar indicates that the number of ChIP-seq reads for that bin and modification is significantly higher or lower than the control, respectively ($P < 0.001$). A '+' or '-' within the bar indicates that a bin is significantly enriched or depleted, respectively, for the histone modification compared to the next lowest expression bin ($P < 0.001$). Error bars shown are the standard error of the mean.

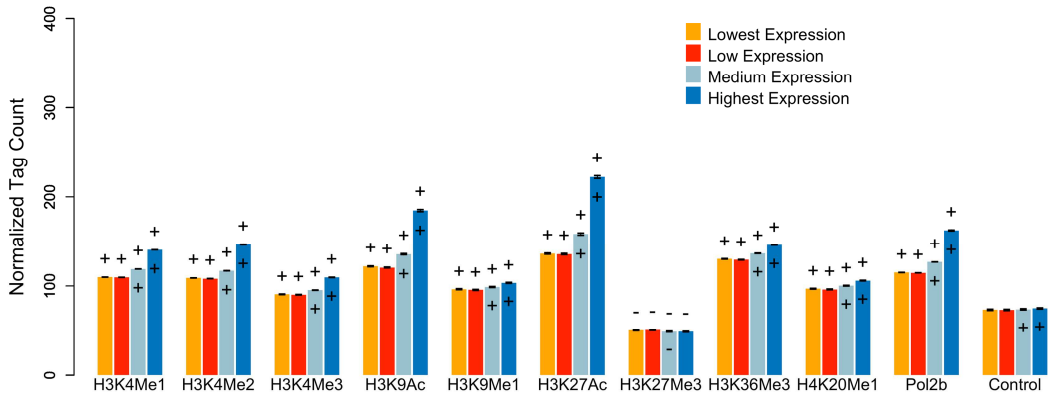


Figure B.2. Enrichment of chromatin modifications and RNA PolII at cis-NAT promoters in NHEK cells using CAGE data from non-polyadenylated nucleus isolates. Cis-NAT promoters were divided into 4 bins based on activity, and the normalized average numbers of ChIP-seq reads from each histone modification +/-5kb of the cis-NAT TSS were calculated for each bin. A '+' or '–' above a bar indicates that the number of ChIP-seq reads for that bin and modification is significantly higher or lower than the control, respectively ($P < 0.001$). A '+' or '-' within the bar indicates that a bin is significantly enriched or depleted, respectively, for the histone modification compared to the next lowest expression bin ($P < 0.001$). Error bars shown are the standard error of the mean.

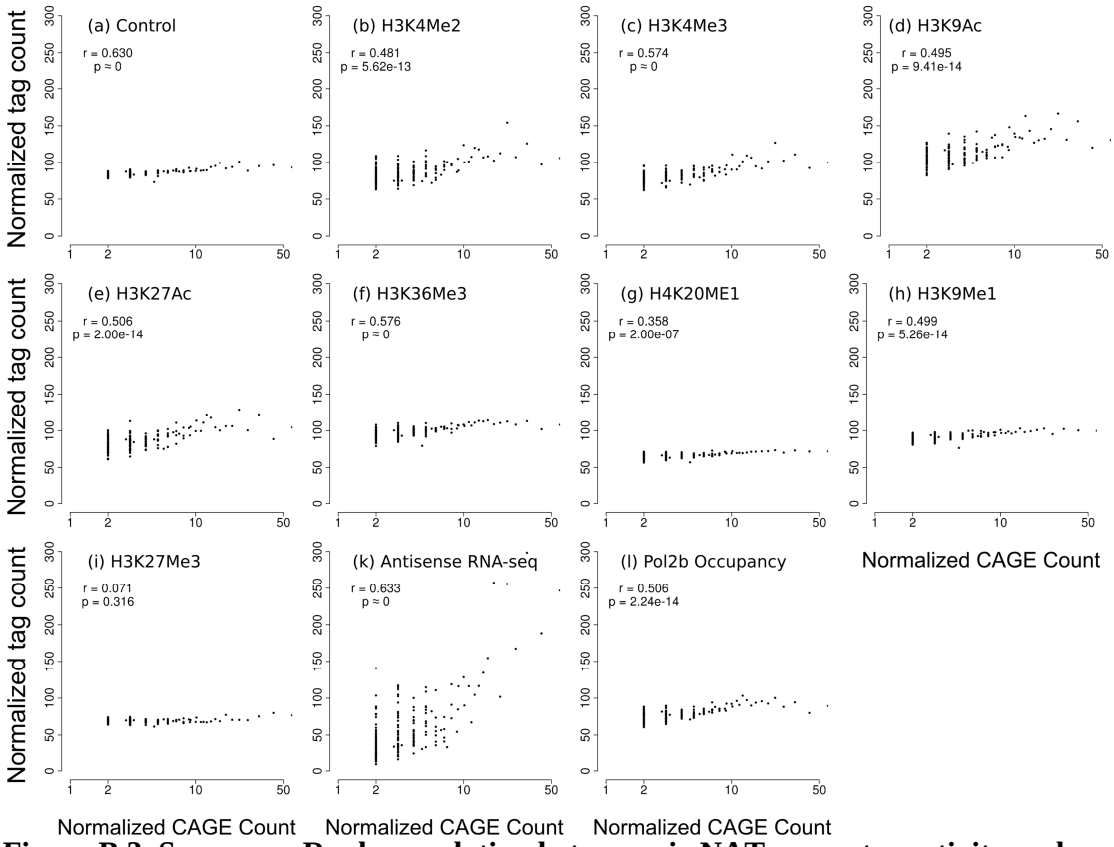


Figure B.3. Spearman Rank correlation between cis-NAT promoter activity and histone modification for K562 nucleus polyadenylated isolates. Cis-NAT promoters were divided into 100 bins based on activity, and the normalized average number of ChIP-seq reads +/-5kb of the cis-NAT TSS were calculated. A Spearman rank correlation was used to determine the relationship between local histone modifications or RNA PolII occupancy and cis-NAT activity.

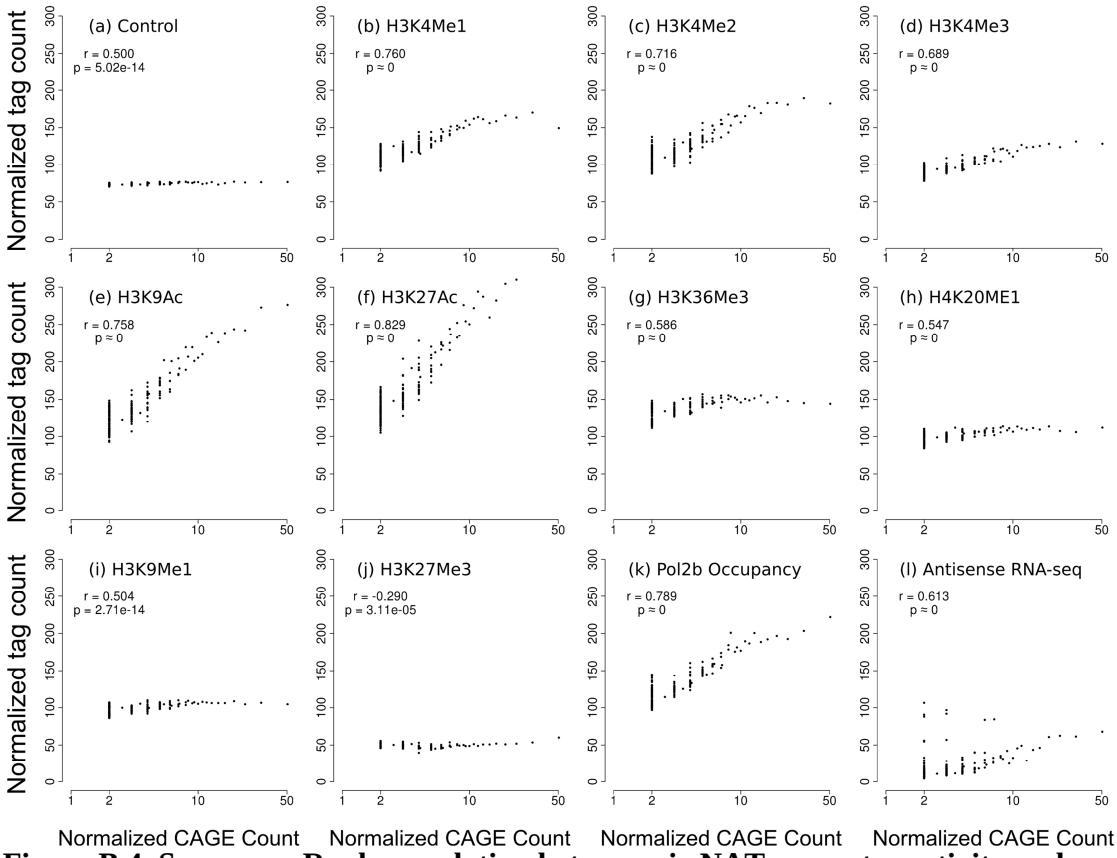


Figure B.4. Spearman Rank correlation between cis-NAT promoter activity and histone modification for NHEK nucleus non-polyadenylated isolates. Cis-NAT promoters were divided into 100 bins based on activity, and the normalized average number of ChIP-seq reads +/-5kb of the cis-NAT TSS were calculated. A Spearman rank correlation was used to determine the relationship between local histone modifications or RNA PolII occupancy and cis-NAT activity.

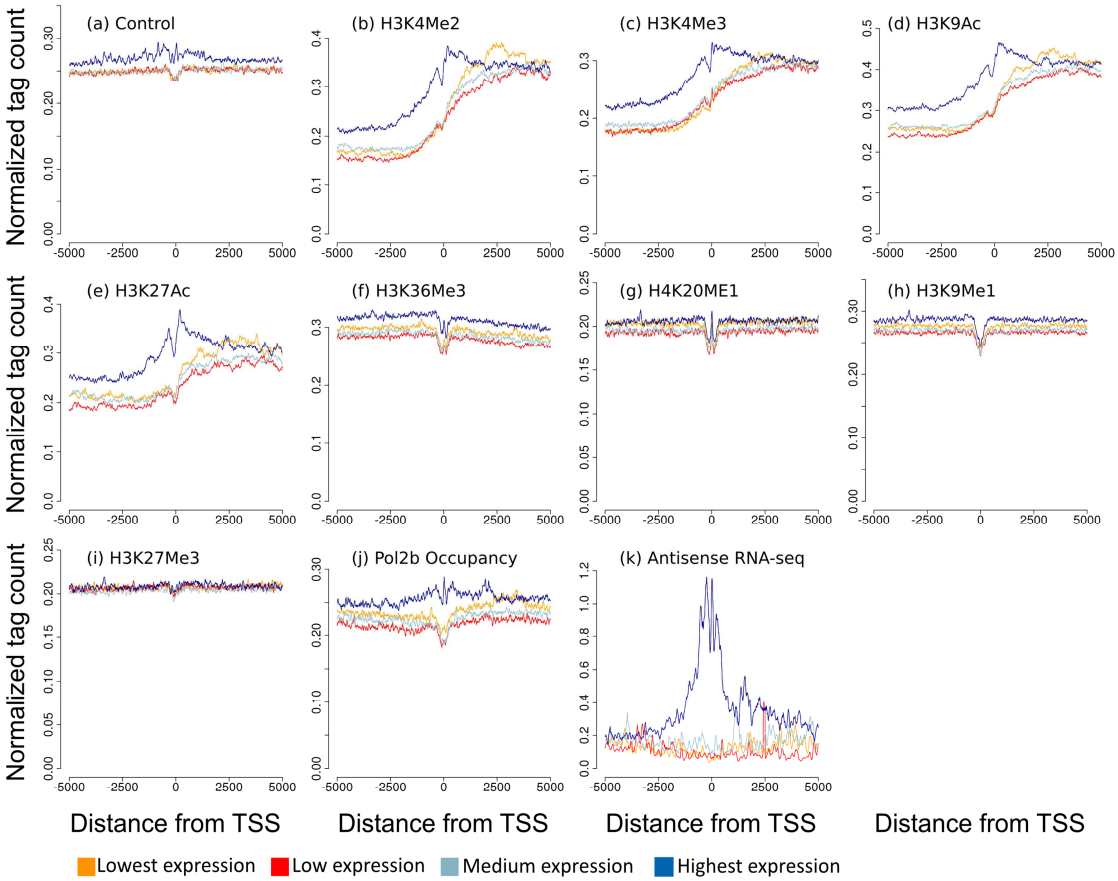


Figure B.5. ChIP-seq read density near cis-NAT TSS in K562 cells using CAGE data from polyadenylated RNA from nuclear isolates. Cis-NAT promoters were divided into 4 bins based on activity, and the normalized average numbers of ChIP-seq reads in 10 base-pair windows within 5kb of the cis-NAT TSS were calculated for each bin.

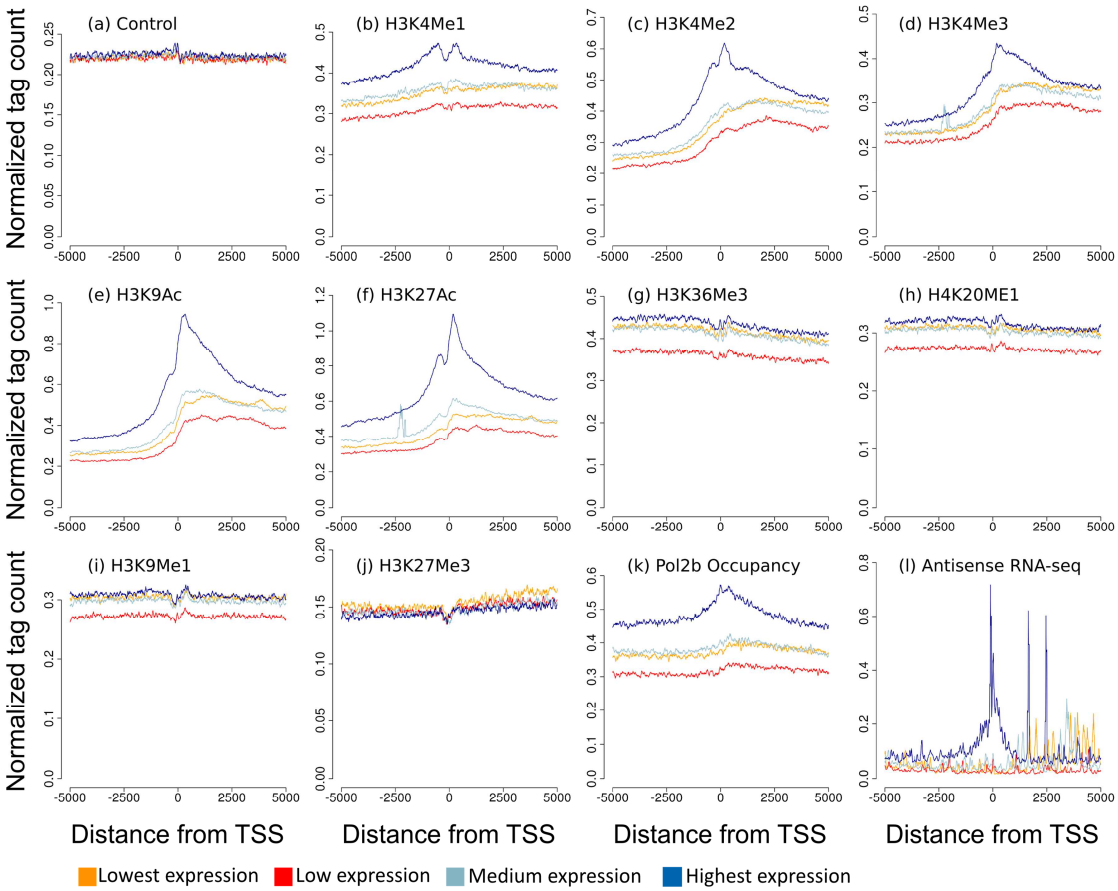


Figure B.6. ChIP-seq read density near cis-NAT TSS in NHEK cells using CAGE data from non-polyadenylated RNA from nuclear isolates. Cis-NAT promoters were divided into 4 bins based on activity, and the normalized average numbers of ChIP-seq reads in 10 base-pair windows within 5kb of the cis-NAT TSS were calculated for each bin.

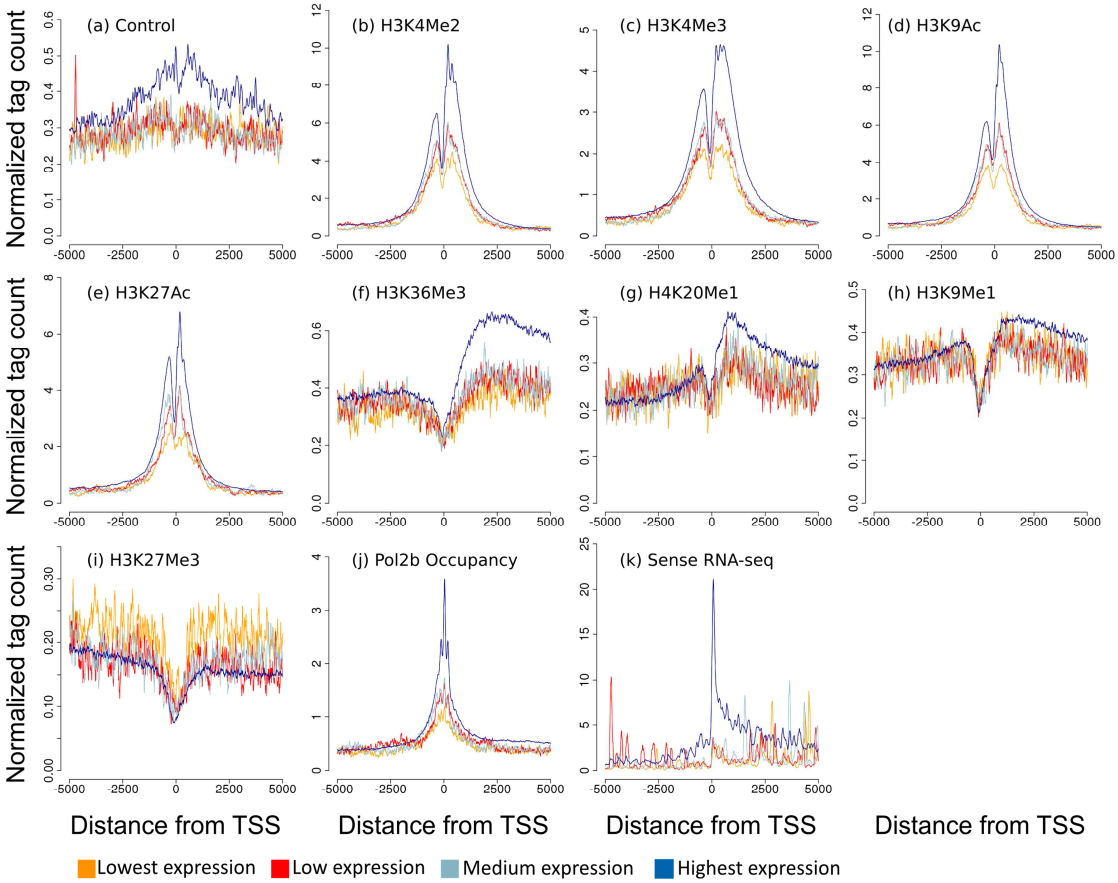


Figure B.7. Chromatin modification environment around genic promoters in K562. Genic promoters were taken from the UCSC genes set, and their activity measured using CAGE data from polyadenylated RNA from nucleus isolates from K562 cells. Genic promoters were divided into the same for bins as cis-NAT promoters for the same data set, and the normalized average numbers of ChIP-seq reads in 10 base-pair windows +/- 5kb of the genic TSS (at position 0) were calculated for each bin.

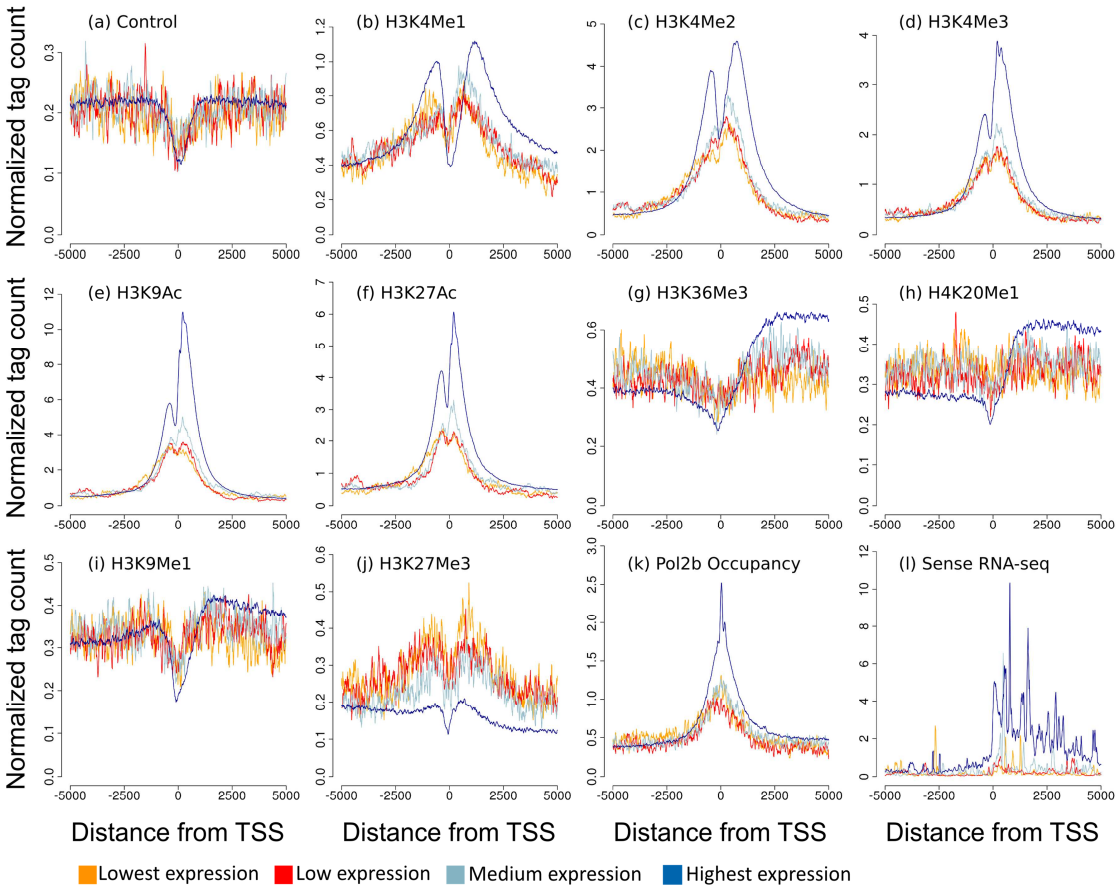


Figure B.8. Chromatin modification environment around genic promoters in NHEK. Genic promoters were taken from the UCSC genes set, and their activity measured using CAGE data from non-polyadenylated RNA from nucleus isolates from NHEK cells. Genic promoters were divided into the same for bins as cis-NAT promoters for the same data set, and the normalized average numbers of ChIP-seq reads in 10 base-pair windows +/- 5kb of the genic TSS (at position 0) were calculated for each bin

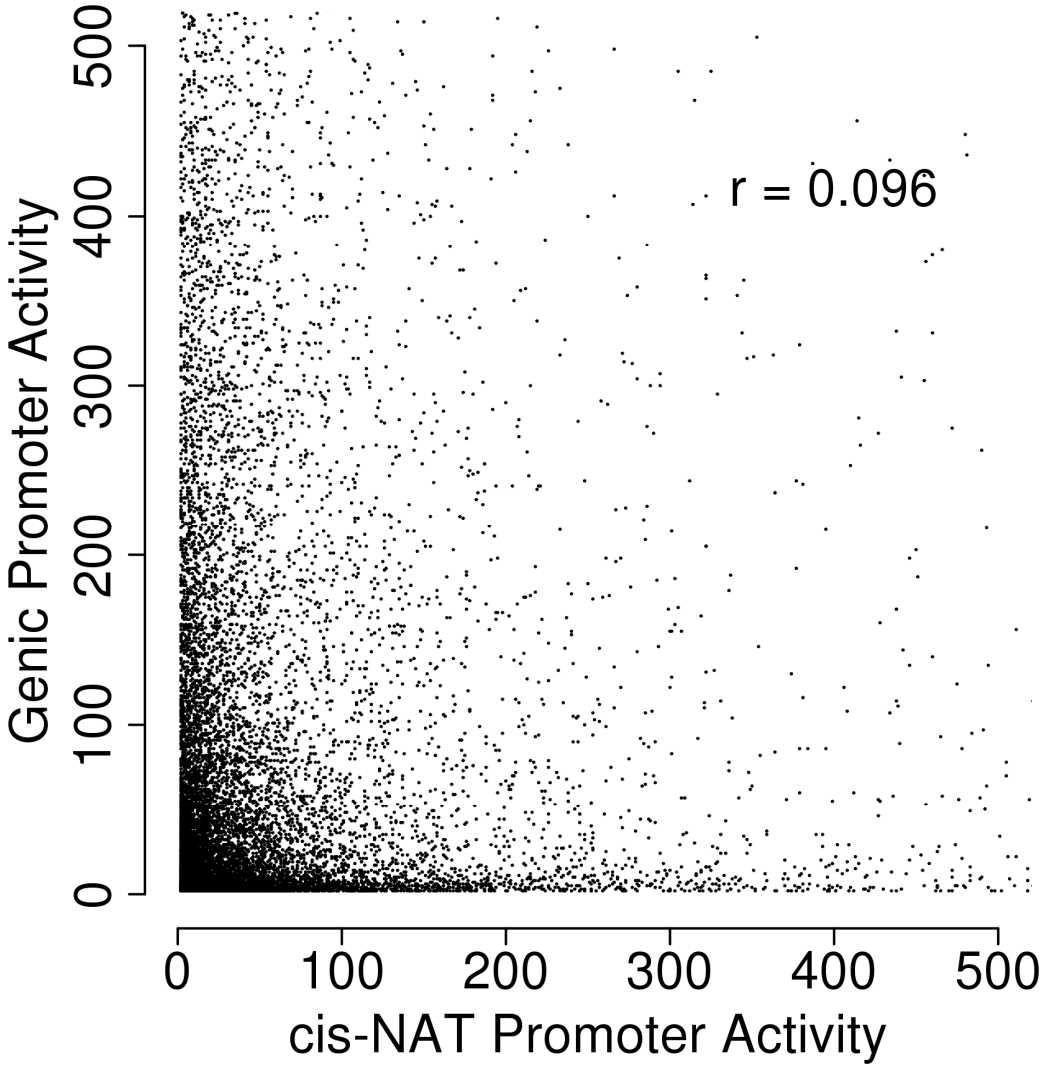


Figure B.9. Correlation between genic and cis-NAT promoter activity in K562. Cis-NAT promoters in the K562 cell type were identified using CAGE data from non-polyadenylated RNA form nucleus isolates. Activity of genic promoters and the sum of corresponding cis-NAT promoter activity was measured by CAGE tag counts. A Spearman rank correlation was used to determine the relationship between total cis-NAT promoter activity and genic promoter activity.

.

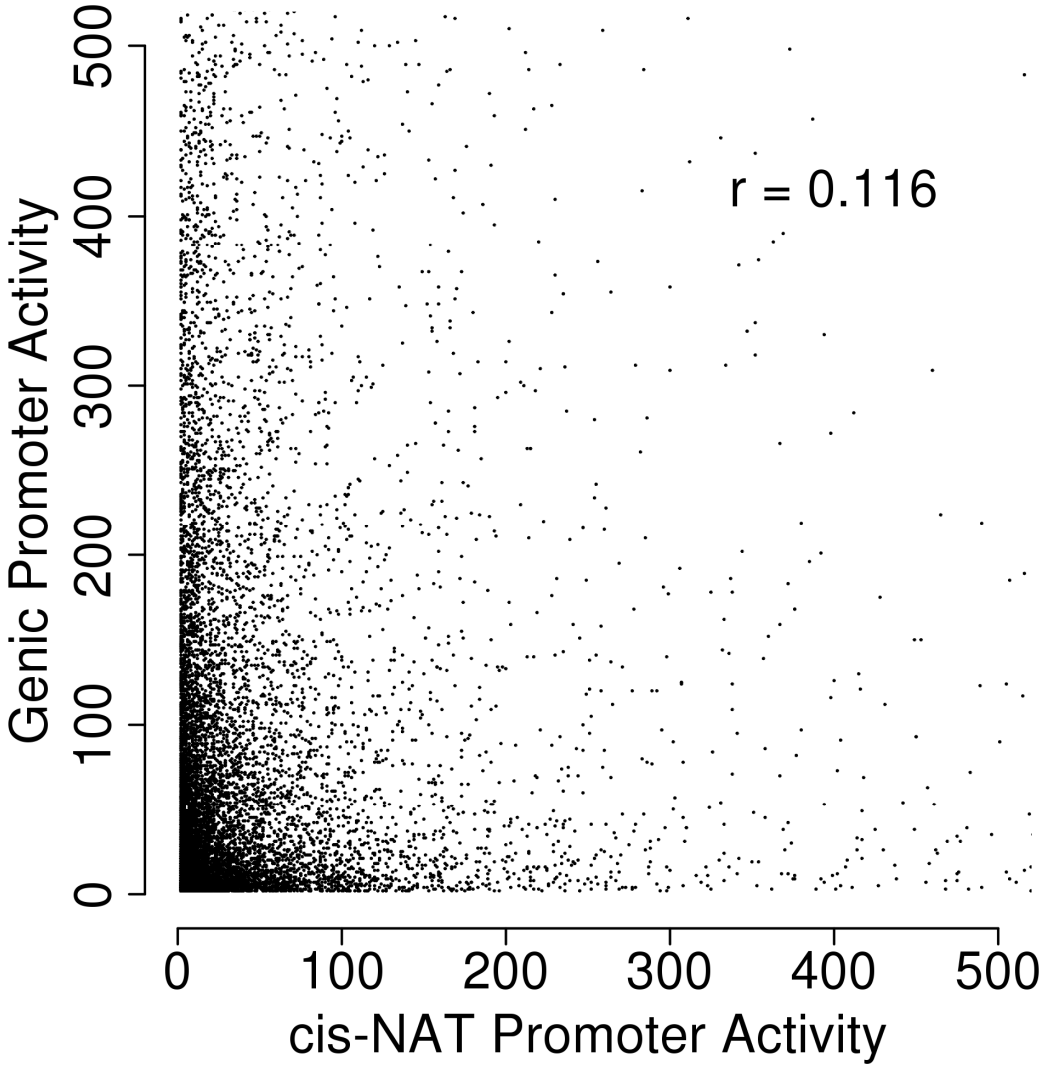


Figure B.10. Correlation between genic and cis-NAT promoter activity in NHEK. Cis-NAT promoters in the NHEK cell type were identified using CAGE data from non-polyadenylated RNA form nucleus isolates. Activity of genic promoters and the sum of corresponding cis-NAT promoter activity was measured by CAGE tag counts. A Spearman rank correlation was used to determine the relationship between total cis-NAT promoter activity and genic promoter activity.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 7

Table C.1. Number of PET tags within TTS clusters, and number of TTS clusters found fore each cell type. PET tag mappings from ENCODE cell types were used to find TTS. Co-locating PET 3' ends were clustered to characterized TTS. Those TTS overlapping TE sequences were found to be TE-TTS.

Cell Type	Sub-Cellular Location	PET Tags in TTS	Non-TE TTS	TE-TTS
GM12878	Nucleus	18,475,428	16,672	2,296
H1HESC	Whole Cell	13,793,627	17,671	1,242
HeLaS3	Nucleus	1,863,548	5,728	407
HepG2	Nucleus	8,934,435	15,883	3,919
HUVEC	Nucleus	3,305,792	18,253	1,247
K562	Nucleus	7,619,273	13,947	2,557
NHEK	Nucleus	17,517,569	15,142	1,126
Prostate	Whole Cell	4,506,631	8,885	794

Table C.2. ChIP-seq reads mapped for each histone modification and cell line. ChIP-seq data from the GM12878 and K562 cell lines were downloaded from the ENCODE repository on the UCSC genome browser. Reads were mapped using bowtie, keeping the best hits with ties broken by quality. Ambiguously mapped reads were resolved using GibbsAM.

Modification	Tags Mapped		
	GM12878	**K562**	**NHEK**
H3K9Ac	12,022,891	17,281,199	12,454,536
H3K27Me3	14,430,662	12,412,831	9,141,036
H3K36Me3	15,195,406	14,950,529	9,182,104

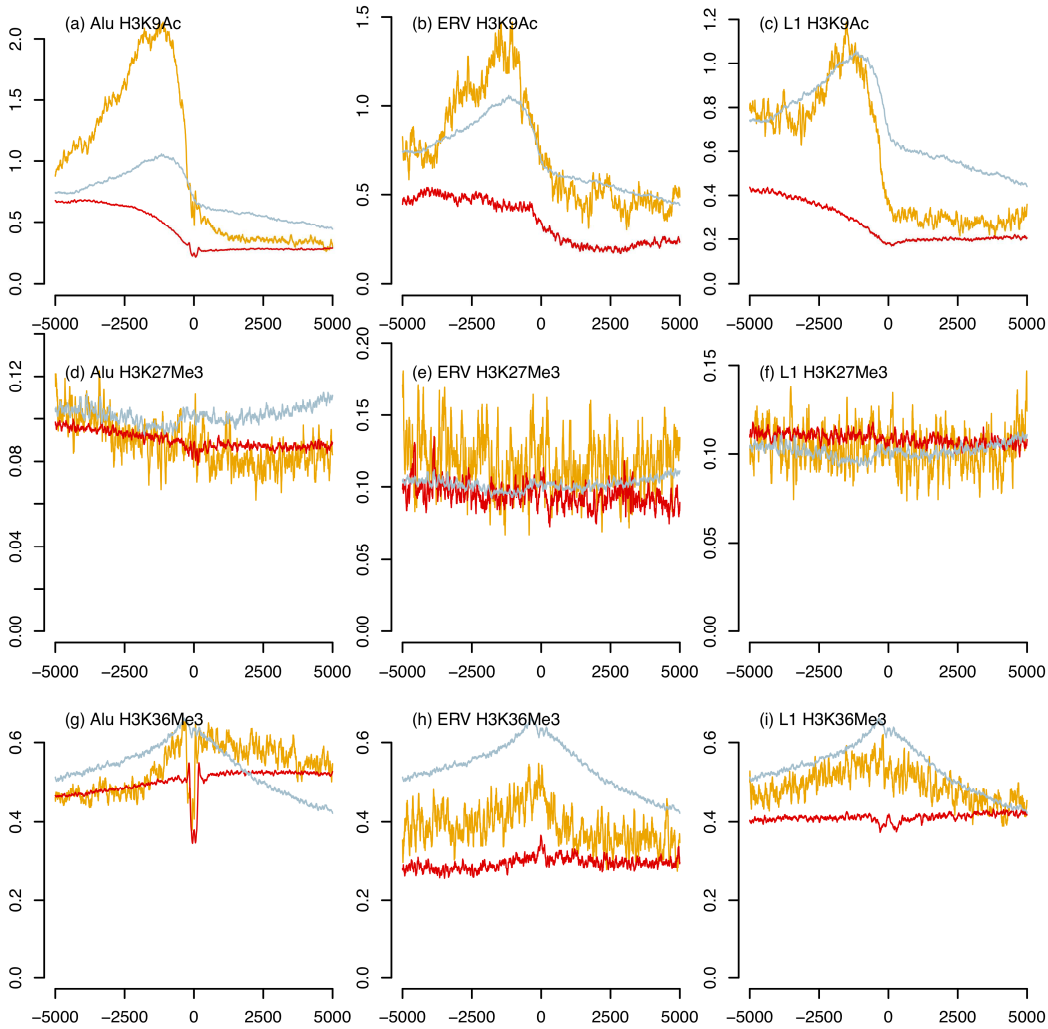


Figure C.1. Enrichment of chromatin modifications at Transcription Termination sites in GM12878. TE-TTS and non TE-TTS were characterized using ENCODE PET data from the GM12878 cell type. Other intragenic TE insertions were defined as those intragenic insertions that do not show a TTS. The average normalized numbers of ChIP-seq tags in 10 base-pair windows +/-5kb of the TTS or insertion were calculated for each set.

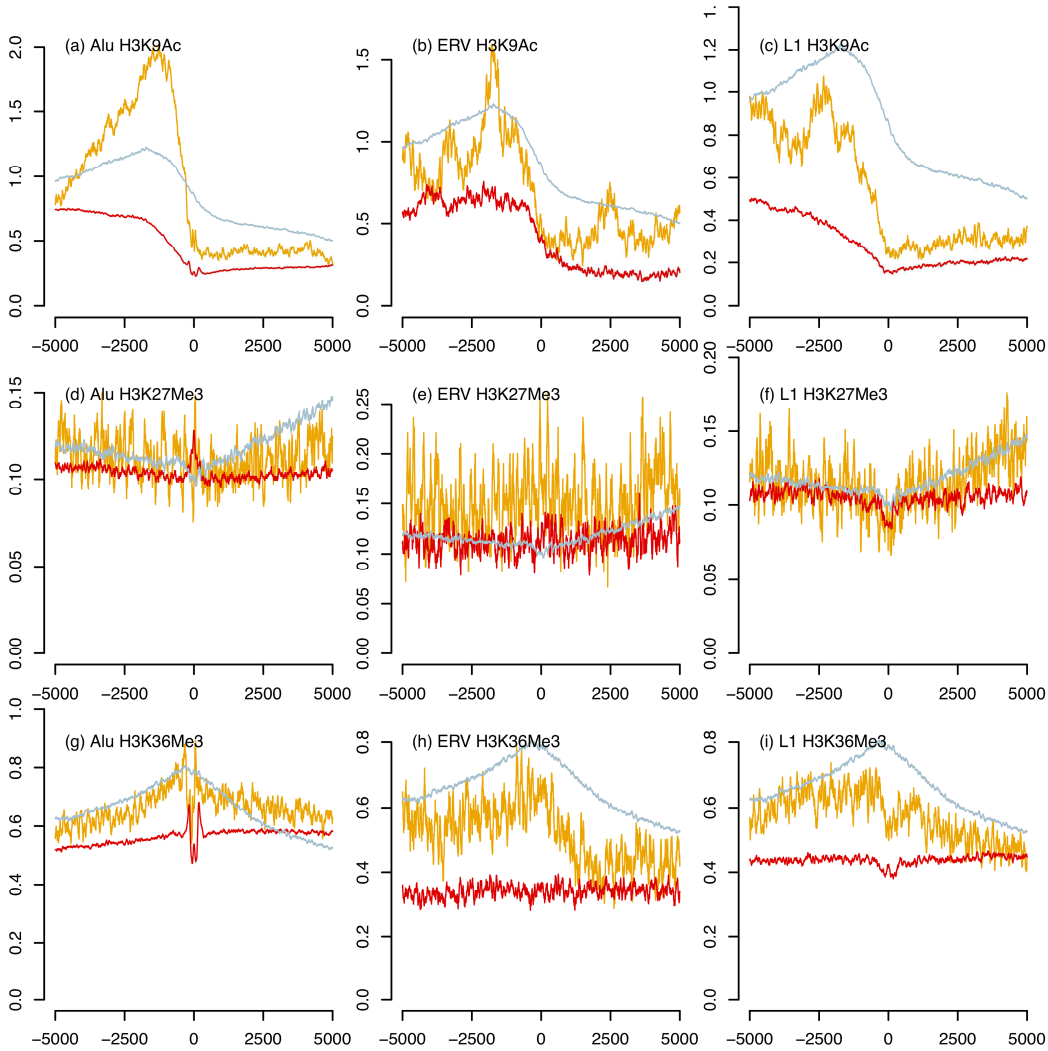


Figure C.2. Enrichment of chromatin modifications at Transcription Termination sites in NHEK. TE-TTS and non TE-TTS were characterized using ENCODE PET data from the NHEK cell type. Other intragenic TE insertions were defined as those intragenic insertions that do not show a TTS. The average normalized numbers of ChIP-seq tags in 10 base-pair windows +/-5kb of the TTS or insertion were calculated for each set.

PUBLICATIONS

Conley, A.B., Miller, W.J. and I.K. Jordan, 2008. Human cis natural antisense transcripts initiated by transposable elements.

Trends Genet. 24: 53-56

Conley, A.B., Piriyapongsa, J. and I.K. Jordan, 2008. Retroviral promoters in the human genome.

Bioinformatics 24: 1563-1567

Conley, A.B. and I.K. Jordan, 2010. Identification of transcription factor binding sites derived from transposable element sequences using ChIP-seq.

Methods Mol. Biol. 674: 225-240

Kislyuk, A.O., Katz, L.S., Agrawal, S., Hagen, M.S., **Conley, A.B.**, Jayaraman, P., Nelakuditi V., Humphrey, J.C., Sammons, S.A., Govil, D., Mair, R.D., Tatti, K.M., Tondella, M.L., Harcourt, B.H., Mayer, L.W. and I.K. Jordan, 2010. A computational genomics pipeline for prokaryotic sequencing projects. 2010 Bioinformatics 26: 1819-1826

Huda, A., Bowen, N.J., **Conley, A.B.** and I.K. Jordan. Epigenetic regulation of transposable element derived human gene promoters. 2011 Gene 475: 39-48

Jordan, I. K., **Conley, A.**, Antonov, I., Arthur, R., Cook, E., Cooper, G., Jones, B., Knipe, K., Lee, K., Liu, X., Mitchell, G., Pande, P., Petit, R., Qin, S., Rajan, V., Sarda, S., Sebastian, A., Tang, S., Thapliyal, R., Varghese, N., Ye, T., Katz, L. S., Wang, X., Rowe, L., Frace, M. and L. Mayer. Genome sequences for five strains of the

emerging pathogen Haemophilus haemolyticus. 2011 J Bacteriol. 193: 5879–5880

Katz, L.S., Humphrey, J.C., **Conley, A.B.**, Nelakuditi, V., Kislyuk, A.O., Agrawal, S., Jayaraman, P., Harcourt, B.H., Olsen- Rasmussen, M.A., Frace, M., Sharma, N.V., Mayer, L.W., and I.K. Jordan. Neisseria Base: a comparative genomics database for Neisseria meningitidis. Database 2011: bar 035

Piriyapongsa, J., Jordan, I.K., **Conley, A.B.**, Ronan, T., and N.R. Smalheiser. Transcription factor binding sites are highly enriched within microRNA precursor sequences 2011 Biol. Direct 6: 61.

Conley, A.B. and I.K. Jordan. Epigenetic regulation of human cis-natural antisense transcripts. 2012 Nucleic Acids Res. 40: 1438–1445.

Lee, K.J., **Conley, A.B.**, Lunyak, V.V., and I.K. Jordan. Do human transposable element small RNAs serve primarily as genome defenders or genome regulators? 2012 Mob Genet Elements. 2: 1–7.

JJingo, D., **Conley, A.B.**, Yi, S.V., Lunyak, V.V., Jordan, I.K. On the presence and role of human gene-body DNA methylation. 2012 Oncotarget. 3:362-74

Conley, A.B. and I.K. Jordan. Endogenous retroviruses and the epigenome. 2012 in press by Springer.

Conley, A.B. and I.K. Jordan. Cell type specific transcription termination by transposable element insertions. 2012 in review in Mobile DNA.

REFERENCES

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

2. Feschotte, C., *Transposable elements and the evolution of regulatory networks.* Nat Rev Genet, 2008. **9**(5): p. 397-405.

3. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

4. Wang, Z., et al., *Combinatorial patterns of histone acetylations and methylations in the human genome.* Nat Genet, 2008. **40**(7): p. 897-903.

5. Smit, A., *RepeatMasker Open-3.0.* 1996-2012.

6. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.

7. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.

8. *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.

9. Quentin, Y., *Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes.* Nucleic Acids Res, 1992. **20**(3): p. 487-93.

10. Quentin, Y., *Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements.* Nucleic Acids Res, 1992. **20**(13): p. 3397-401.

11. Ostertag, E.M., et al., *SVA elements are nonautonomous retrotransposons that cause disease in humans.* Am J Hum Genet, 2003. **73**(6): p. 1444-51.

12. Silva, J.C., et al., *Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes.* Genet Res, 2003. **82**(1): p. 1-18.

13. Nur, I., E. Pascale, and A.V. Furano, *The left end of rat L1 (L1Rn, long interspersed repeated) DNA which is a CpG island can function as a promoter.* Nucleic Acids Res, 1988. **16**(19): p. 9233-51.

14. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-74.

15. Schibler, U., et al., *Two promoters of different strengths control the transcription of the mouse alpha-amylase gene Amy-1a in the parotid gland and the liver.* Cell, 1983. **33**(2): p. 501-8.

16. Chretien, S., et al., *Alternative transcription and splicing of the human porphobilinogen deaminase gene result either in tissue-specific or in housekeeping expression.* Proc Natl Acad Sci U S A, 1988. **85**(1): p. 6-10.

17. Stapleton, G., M.P. Somma, and P. Lavia, *Cell type-specific interactions of transcription factors with a housekeeping promoter in vivo.* Nucleic Acids Res, 1993. **21**(10): p. 2465-71.

18. Cohen, C.J., et al., *Placenta-specific expression of the interleukin-2 (IL-2) receptor beta subunit from an endogenous retroviral promoter.* J Biol Chem, 2011. **286**(41): p. 35543-52.

19. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.

20. Mager, D.L., *Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H.* Virology, 1989. **173**(2): p. 591-9.

21. Mager, D.L., et al., *Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3).* Genomics, 1999. **59**(3): p. 255-63.

22. Nishihara, H., A.F. Smit, and N. Okada, *Functional noncoding sequences derived from SINEs in the mammalian genome.* Genome Res, 2006. **16**(7): p. 864-74.

23. Sasaki, T., et al., *Possible involvement of SINEs in mammalian-specific brain formation.* Proc Natl Acad Sci U S A, 2008. **105**(11): p. 4220-5.

24. Schmid, C.D. and P. Bucher, *MER41 repeat sequences contain inducible STAT1 binding sites.* PLoS One, 2010. **5**(7): p. e11425.

25. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

26. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

27. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.

28. Rosenbloom, K.R., et al., *ENCODE whole-genome data in the UCSC Genome Browser.* Nucleic Acids Res, 2010. **38**(Database issue): p. D620-5.

29. Raney, B.J., et al., *ENCODE whole-genome data in the UCSC genome browser (2011 update).* Nucleic Acids Res, 2011. **39**(Database issue): p. D871-5.

30. Shiraki, T., et al., *Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15776-81.

31. Kodzius, R., et al., *CAGE: cap analysis of gene expression.* Nat Methods, 2006. **3**(3): p. 211-22.

32. Faulkner, G.J., et al., *The regulated retrotransposon transcriptome of mammalian cells.* Nat Genet, 2009. **41**(5): p. 563-71.

33. Yelin, R., et al., *Widespread occurrence of antisense transcription in the human genome.* Nat Biotechnol, 2003. **21**(4): p. 379-86.

34. Lapidot, M. and Y. Pilpel, *Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms.* EMBO Rep, 2006. **7**(12): p. 1216-22.

35. Faghihi, M.A. and C. Wahlestedt, *Regulatory roles of natural antisense transcripts.* Nat Rev Mol Cell Biol, 2009. **10**(9): p. 637-43.

36. Modarresi, F., et al., *Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation.* Nat Biotechnol, 2012. **30**(5): p. 453-9.

37. Lutz, C.S., *Alternative polyadenylation: a twist on mRNA 3' end formation.* ACS Chem Biol, 2008. **3**(10): p. 609-17.

38. Lutz, C.S. and A. Moreira, *Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression.* Wiley Interdiscip Rev RNA, 2011. **2**(1): p. 22-31.

39. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.

40. Ji, Z. and B. Tian, *Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types.* PLoS One, 2009. **4**(12): p. e8419.

41. Ji, Z., et al., *Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development.* Proc Natl Acad Sci U S A, 2009. **106**(17): p. 7028-33.

42. Chen, J., et al., *Over 20% of human transcripts might form sense-antisense pairs.* Nucleic Acids Res, 2004. **32**(16): p. 4812-20.

43. Lehner, B., et al., *Antisense transcripts in the human genome.* Trends Genet, 2002. **18**(2): p. 63-5.

44. Osato, N., et al., *Transcriptional interferences in cis natural antisense transcripts of humans and mice.* Genetics, 2007. **176**(2): p. 1299-306.

45. Fire, A., *RNA-triggered gene silencing.* Trends Genet, 1999. **15**(9): p. 358-63.

46. Matzke, M.A., M.F. Mette, and A.J. Matzke, *Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates.* Plant Mol Biol, 2000. **43**(2-3): p. 401-15.

47. Slotkin, R.K., M. Freeling, and D. Lisch, *Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication.* Nat Genet, 2005. **37**(6): p. 641-4.

48. Vastenhouw, N.L. and R.H. Plasterk, *RNAi protects the Caenorhabditis elegans germline against transposition.* Trends Genet, 2004. **20**(7): p. 314-9.

49. Piriyapongsa, J. and I.K. Jordan, *A family of human microRNA genes from miniature inverted-repeat transposable elements.* PLoS One, 2007. **2**(2): p. e203.

50. Smalheiser, N.R. and V.I. Torvik, *Mammalian microRNAs derived from genomic repeats.* Trends Genet, 2005. **21**(6): p. 322-6.

51. Vagin, V.V., et al., *A distinct small RNA pathway silences selfish genetic elements in the germline.* Science, 2006. **313**(5785): p. 320-4.

52. Brennecke, J., et al., *Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila.* Cell, 2007. **128**(6): p. 1089-103.

53. Thornburg, B.G., V. Gotea, and W. Makalowski, *Transposable elements as a significant source of transcription regulating signals.* Gene, 2006. **365**: p. 104-10.

54. Carninci, P., et al., *Genome-wide analysis of mammalian promoter architecture and evolution.* Nat Genet, 2006. **38**(6): p. 626-35.

55. Karolchik, D., et al., *The UCSC Genome Browser Database.* Nucleic Acids Res, 2003. **31**(1): p. 51-4.

56. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.

57. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.

58. Jurka, J., J. Walichiewicz, and A. Milosavljevic, *Prototypic sequences for human repetitive DNA.* J Mol Evol, 1992. **35**(4): p. 286-91.

59. Jurka, J., *Repbase update: a database and an electronic journal of repetitive elements.* Trends Genet, 2000. **16**(9): p. 418-20.

60. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.

61. Kapitonov, V. and J. Jurka, *The age of Alu subfamilies.* J Mol Evol, 1996. **42**(1): p. 59-65.

62. Jukes, T.H.C., C.R. , *Evolution of protein molecules,* in *Mammalian protein metabolism*, H.D. Munro, Editor 1969, Academic Press: New York.

63. Schwartz, S., et al., *Human-mouse alignments with BLASTZ.* Genome Res, 2003. **13**(1): p. 103-7.

64. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.

65. Okazaki, Y., et al., *Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.* Nature, 2002. **420**(6915): p. 563-73.

66. Britten, R.J., *DNA sequence insertion and evolutionary variation in gene regulation.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9374-7.

67. Britten, R.J., *Mobile elements inserted in the distant past have taken on important functions.* Gene, 1997. **205**(1-2): p. 177-82.

68. Jordan, I.K., et al., *Origin of a substantial fraction of human regulatory sequences from transposable elements.* Trends Genet, 2003. **19**(2): p. 68-72.

69. Marino-Ramirez, L. and I.K. Jordan, *Transposable element derived DNaseI-hypersensitive sites in the human genome.* Biol Direct, 2006. **1**: p. 20.

70. van de Lagemaat, L.N., et al., *Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions.* Trends Genet, 2003. **19**(10): p. 530-6.

71. Bock, M. and J.P. Stoye, *Endogenous retroviruses and the human germline.* Curr Opin Genet Dev, 2000. **10**(6): p. 651-5.

72. Bromham, L., *The human zoo: endogenous retroviruses in the human genome.* Trends in Ecology & Evolution, 2002. **17**(2): p. 91-97.

73. Sverdlov, E.D., *Retroviruses and primate evolution.* Bioessays, 2000. **22**(2): p. 161-71.

74. Bannert, N. and R. Kurth, *Retroelements and the human genome: new perspectives on an old relation.* Proc Natl Acad Sci U S A, 2004. **101 Suppl 2**: p. 14572-9.

75. Samuelson, L.C., et al., *Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution.* Mol Cell Biol, 1990. **10**(6): p. 2513-20.

76. Dunn, C.A., P. Medstrand, and D.L. Mager, *An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12841-6.

77. Dunn, C.A., et al., *Transcription of two human genes from a bidirectional endogenous retrovirus promoter.* Gene, 2006. **366**(2): p. 335-42.

78. Romanish, M.T., et al., *Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution.* PLoS Genet, 2007. **3**(1): p. e10.

79. King, M.C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-16.

80. Ng, P., et al., *Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.* Nat Methods, 2005. **2**(2): p. 105-11.

81. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.

82. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.

83. Blanchette, M., et al., *Aligning multiple genomic sequences with the threaded blockset aligner.* Genome Res, 2004. **14**(4): p. 708-15.

84. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

85. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

86. Safran, M., et al., *GeneCards 2002: towards a complete, object-oriented, human gene compendium.* Bioinformatics, 2002. **18**(11): p. 1542-3.

87. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.

88. Kodym, R., P. Calkins, and M. Story, *The cloning and characterization of a new stress response protein. A mammalian member of a family of theta class glutathione s-transferase-like proteins.* J Biol Chem, 1999. **274**(8): p. 5131-7.

89. Kolsch, H., et al., *Polymorphisms in glutathione S-transferase omega-1 and AD, vascular dementia, and stroke.* Neurology, 2004. **63**(12): p. 2255-60.

90. Li, Y.J., et al., *Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease.* Hum Mol Genet, 2003. **12**(24): p. 3259-67.

91. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements.* Nat Rev Genet, 2007. **8**(12): p. 973-82.

92. Doolittle, W.F. and C. Sapienza, *Selfish genes, the phenotype paradigm and genome evolution.* Nature, 1980. **284**(5757): p. 601-3.

93. Orgel, L.E. and F.H. Crick, *Selfish DNA: the ultimate parasite.* Nature, 1980. **284**(5757): p. 604-7.

94. Gould, S.J.V., E.S., *Exaptation; a missing term in the science of form.* Paleobiology, 1982. **8**(1): p. 4-15.

95. Jordan, I.K., *Evolutionary tinkering with transposable elements.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 7941-2.

96. Kidwell, M.G. and D.R. Lisch, *Perspective: transposable elements, parasitic DNA, and genome evolution.* Evolution, 2001. **55**(1): p. 1-24.

97. Cohen, C.J., W.M. Lock, and D.L. Mager, *Endogenous retroviral LTRs as promoters for human genes: a critical assessment.* Gene, 2009. **448**(2): p. 105-14.

98. Conley, A.B., J. Piriyapongsa, and I.K. Jordan, *Retroviral promoters in the human genome.* Bioinformatics, 2008. **24**(14): p. 1563-7.

99. Zemojtel, T., et al., *Methylation and deamination of CpGs generate p53-binding sites on a genomic scale.* Trends Genet, 2009. **25**(2): p. 63-6.

100. Wang, J., et al., *A c-Myc regulatory subnetwork from human transposable element sequences.* Mol Biosyst, 2009. **5**(12): p. 1831-9.

101. Marino-Ramirez, L., et al., *Transposable elements donate lineage-specific regulatory sequences to host genomes.* Cytogenet Genome Res, 2005. **110**(1-4): p. 333-41.

102. Zhang, Z. and M. Gerstein, *Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.* J Biol, 2003. **2**(2): p. 11.

103. Lowe, C.B., G. Bejerano, and D. Haussler, *Thousands of human mobile element fragments undergo strong purifying selection near developmental genes.* Proc Natl Acad Sci U S A, 2007. **104**(19): p. 8005-10.

104. Santangelo, A.M., et al., *Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene.* PLoS Genet, 2007. **3**(10): p. 1813-26.

105. Hirakawa, M., et al., *Characterization and evolutionary landscape of AmnSINE1 in Amniota genomes.* Gene, 2009. **441**(1-2): p. 100-10.

106. Smith, A.M., et al., *A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity.* Genome Res, 2008. **18**(9): p. 1422-32.

107. Pang, K.C., M.C. Frith, and J.S. Mattick, *Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function.* Trends Genet, 2006. **22**(1): p. 1-5.

108. Johnson, R., et al., *Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication.* Nucleic Acids Res, 2006. **34**(14): p. 3862-77.

109. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome Res, 2008. **18**(11): p. 1752-62.

110. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

111. Hashimoto, T., et al., *Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite.* Bioinformatics, 2009. **25**(19): p. 2613-4.

112. Kuhn, R.M., et al., *The UCSC Genome Browser Database: update 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D755-61.

113. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

114. Burrows, M.W., D.J., *A block-sorting lossless data compression algorithm.* Digital Systems Research Center, 1994.

115. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.

116. Faulkner, G.J., et al., *A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE.* Genomics, 2008. **91**(3): p. 281-8.

117. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.

118. Jothi, R., et al., *Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.* Nucleic Acids Res, 2008. **36**(16): p. 5221-31.

119. Gaszner, M. and G. Felsenfeld, *Insulators: exploiting transcriptional and epigenetic mechanisms.* Nat Rev Genet, 2006. **7**(9): p. 703-13.

120. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-45.

121. Frith, M.C., et al., *Detection of functional DNA motifs via statistical over-representation.* Nucleic Acids Res, 2004. **32**(4): p. 1372-81.

122. Ondov, B.D., et al., *Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications.* Bioinformatics, 2008. **24**(23): p. 2776-7.

123. Bailey, T.L. and M. Gribskov, *Combining evidence using p-values: application to sequence homology searches.* Bioinformatics, 1998. **14**(1): p. 48-54.

124. Conley, A.B., W.J. Miller, and I.K. Jordan, *Human cis natural antisense transcripts initiated by transposable elements.* Trends Genet, 2008. **24**(2): p. 53-6.

125. Tam, O.H., et al., *Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes.* Nature, 2008. **453**(7194): p. 534-8.

126. Watanabe, T., et al., *Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes.* Nature, 2008. **453**(7194): p. 539-43.

127. Struhl, K., *Transcriptional noise and the fidelity of initiation by RNA polymerase II.* Nat Struct Mol Biol, 2007. **14**(2): p. 103-5.

128. Wang, J., et al., *Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs.* Nature, 2004. **431**(7010): p. 1 p following 757; discussion following 757.

129. Werner, A. and A. Berdal, *Natural antisense transcripts: sound or silence?* Physiol Genomics, 2005. **23**(2): p. 125-31.

130. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.

131. Trinklein, N.D., et al., *Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome.* Genome Res, 2007. **17**(6): p. 720-31.

132. Hon, G., W. Wang, and B. Ren, *Discovery and annotation of functional chromatin signatures in the human genome.* PLoS Comput Biol, 2009. **5**(11): p. e1000566.

133. Rhead, B., et al., *The UCSC Genome Browser database: update 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D613-9.

134. Wang, J., et al., *A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags.* Bioinformatics, 2010. **26**(20): p. 2501-8.

135. Tan, P.N., Steinbach, M. and Kumar, V., *Introduction to Data Mining*2005, Boston: Addison-Wesley.

136. Sokal, R.R.a.R., J.F., *Biometry: The Principles and Practice of Statistics in Biological Research.* 1981, San Francisco: W. H. Freeman.

137. Mercer, T.R., et al., *Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome.* Genome Res, 2010. **20**(12): p. 1639-50.

138. Lower, R., J. Lower, and R. Kurth, *The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences.* Proc Natl Acad Sci U S A, 1996. **93**(11): p. 5177-84.

139. Xiong, Y. and T.H. Eickbush, *Origin and evolution of retroelements based upon their reverse transcriptase sequences.* EMBO J, 1990. **9**(10): p. 3353-62.

140. Xiong, Y. and T.H. Eickbush, *Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns.* Mol Biol Evol, 1988. **5**(6): p. 675-90.

141. Doolittle, R.F., et al., *Origins and evolutionary relationships of retroviruses.* Q Rev Biol, 1989. **64**(1): p. 1-30.

142. Maksakova, I.A., et al., *Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line.* PLoS Genet, 2006. **2**(1): p. e2.

143. Wu, J., et al., *Autoimmune disease in mice due to integration of an endogenous retrovirus in an apoptosis gene.* J Exp Med, 1993. **178**(2): p. 461-8.

144. Li, J., et al., *Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance.* Genome Res, 2012.

145. Levin, H.L. and J.V. Moran, *Dynamic interactions between transposable elements and their hosts.* Nat Rev Genet, 2011. **12**(9): p. 615-27.

146. Lippman, Z., et al., *Role of transposable elements in heterochromatin and epigenetic control.* Nature, 2004. **430**(6998): p. 471-6.

147. Leung, D.C. and M.C. Lorincz, *Silencing of endogenous retroviruses: when and why do histone marks predominate?* Trends Biochem Sci, 2011.

148. Martens, J.H., et al., *The profile of repeat-associated histone lysine methylation states in the mouse epigenome.* EMBO J, 2005. **24**(4): p. 800-12.

149. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

150. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.

151. Huda, A., L. Marino-Ramirez, and I.K. Jordan, *Epigenetic histone modifications of human transposable elements: genome defense versus exaptation.* Mob DNA, 2010. **1**(1): p. 2.

152. Brosius, J. and S.J. Gould, *On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA".* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10706-10.

153. Hollister, J.D. and B.S. Gaut, *Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression.* Genome Res, 2009. **19**(8): p. 1419-28.

154. Rebollo, R., et al., *Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms.* PLoS Genet, 2011. **7**(9): p. e1002301.

155. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.

156. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.

157. Michaud, E.J., et al., *Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage.* Genes Dev, 1994. **8**(12): p. 1463-72.

158. Morgan, H.D., et al., *Epigenetic inheritance at the agouti locus in the mouse.* Nat Genet, 1999. **23**(3): p. 314-8.

159. Peaston, A.E., et al., *Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos.* Dev Cell, 2004. **7**(4): p. 597-606.

160. Huda, A., et al., *Epigenetic regulation of transposable element derived human gene promoters.* Gene, 2011. **475**(1): p. 39-48.

161. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.

162. Huda, A., et al., *Prediction of transposable element derived enhancers using chromatin modification profiles.* PLoS One, 2011. **6**(11): p. e27513.

163. McClintock, B., *Mutable Loci in Maize.* Carnegie Institute of Washingto Yearbook, 1948. **47**: p. 155-169.

164. McClintock, B., *The significance of responses of the genome to challenge.* Science, 1984. **226**(4676): p. 792-801.

165. Feil, R. and M.F. Fraga, *Epigenetics and the environment: emerging patterns and implications.* Nat Rev Genet, 2011. **13**(2): p. 97-109.

166. Cropley, J.E., et al., *Germ-line epigenetic modification of the murine A vy allele by nutritional supplementation.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17308-12.

167. Smit, A.F., *Interspersed repeats and other mementos of transposable elements in mammalian genomes.* Curr Opin Genet Dev, 1999. **9**(6): p. 657-63.

168. Ono, M., *Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes.* J Virol, 1986. **58**(3): p. 937-44.

169. Goodchild, N.L., D.A. Wilkinson, and D.L. Mager, *A human endogenous long terminal repeat provides a polyadenylation signal to a novel, alternatively spliced transcript in normal placenta.* Gene, 1992. **121**(2): p. 287-94.

170. Ustyugova, S.V., Y.B. Lebedev, and E.D. Sverdlov, *Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts.* Genetica, 2006. **128**(1-3): p. 261-72.

171. Medstrand, P., L.N. van de Lagemaat, and D.L. Mager, *Retroelement distributions in the human genome: variations associated with age and proximity to genes.* Genome Res, 2002. **12**(10): p. 1483-95.

172. Lee, J.Y., Z. Ji, and B. Tian, *Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes.* Nucleic Acids Res, 2008. **36**(17): p. 5581-90.

173. Chen, C., T. Ara, and D. Gautheret, *Using Alu elements as polyadenylation sites: A case of retroposon exaptation.* Mol Biol Evol, 2009. **26**(2): p. 327-34.

174. Fujita, P.A., et al., *The UCSC Genome Browser database: update 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D876-82.

175. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq.* RNA, 2011. **17**(4): p. 761-72.

176. Eberle, A.B., et al., *Posttranscriptional gene regulation by spatial rearrangement of the 3' untranslated region.* PLoS Biol, 2008. **6**(4): p. e92.

177. Yepiskoposyan, H., et al., *Autoregulation of the nonsense-mediated mRNA decay pathway in human cells.* RNA, 2011. **17**(12): p. 2108-18.

178. Roy-Engel, A.M., et al., *Human retroelements may introduce intragenic polyadenylation signals.* Cytogenet Genome Res, 2005. **110**(1-4): p. 365-71.

179. *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.

180. Vorlova, S., et al., *Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation.* Mol Cell, 2011. **43**(6): p. 927-39.

181. Jenal, M., et al., *The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites.* Cell, 2012. **149**(3): p. 538-53.

182. Ebert, M.S., J.R. Neilson, and P.A. Sharp, *MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells.* Nat Methods, 2007. **4**(9): p. 721-6.

183. Ebert, M.S. and P.A. Sharp, *MicroRNA sponges: progress and possibilities.* RNA, 2010. **16**(11): p. 2043-50.

184. Carninci, P., et al., *High-efficiency full-length cDNA cloning by biotinylated CAP trapper.* Genomics, 1996. **37**(3): p. 327-36.

185. Jordan, I.K., et al., *Conservation and coevolution in the scale-free human gene coexpression network.* Mol Biol Evol, 2004. **21**(11): p. 2058-70.

186. Jordan, I.K., L. Marino-Ramirez, and E.V. Koonin, *Evolutionary significance of gene expression divergence.* Gene, 2005. **345**(1): p. 119-26.

187. Yanai, I., et al., *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.* Bioinformatics, 2005. **21**(5): p. 650-9.

188. Sturn, A., J. Quackenbush, and Z. Trajanoski, *Genesis: cluster analysis of microarray data.* Bioinformatics, 2002. **18**(1): p. 207-8.

189. Zhang, B., et al., *GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.* BMC Bioinformatics, 2004. **5**: p. 16.